

Faktorenanalyse

1. Schritt: Inhalte der Faktorenanalyse

- Verfahren zur Summation oder Reduktion von Daten
- R-Analyse: Untersuchung der Beziehungen von Variablen (das übliche)
Q-Analyse: Untersuchung der Beziehungen von Personen
 - vgl. Slicing the data box
 - Q-Analyse bestimmt Personengruppen mit gemeinsamem Beziehungsmuster, während die Clusteranalyse Personengruppen mit der größten Nähe bestimmt.
- Summation: eine Skala wird gebildet, die einen Gesamtwert der einbezogenen Variablen darstellt
- Datenreduktion: Auswahl einiger repräsentativer Variablen, um eine sparsame Darstellung der Informationen zu erhalten.
- Auswahl der Input-Variablen: begründet durch theoretische Überlegungen; durch die Wahl der Variablen werden die potentiell identifizierbaren Dimensionen vorbestimmt!

2. Schritt: Design der Faktorenanalyse

- Berechnung einer Korrelationsmatrix als Input für die Analyse
- die Variablen sollten metrisch sein oder als Dummies kodiert werden (wenn alle Variablen dummykodiert sind, ergibt sich die Bool'sche FA)
- die Zahl der Variablen sollte nicht zu groß sein, jedoch sollten pro Faktor etwa 5 Variablen in die Analyse aufgenommen werden
- Stichprobe sollte mindestens $N=100$ sein, bzw. mindestens 5 Beobachtungen pro Variable (ansonsten overfitting, wie immer)

3. Schritt: Überprüfung der Annahmen

- Abweichungen von Normalverteilung, Homoskedastizität und Linearität sind nur insofern von Bedeutung, als daß sie die Korrelationen der Variablen beeinflussen
- Normalverteilung ist nur dann von Bedeutung, wenn Signifikanztests für die Faktoren berechnet werden, was sehr selten ist.
- Multikollinearität ist erwünscht!
- wenn wenige der Interkorrelationen größer als .30 sind, ist FA nicht angemessen
- bei Betrachtung der Partialkorrelationen sollten einige Werte sehr gering sein, das deutet auf Faktoren hin („anti-image matrix“ im SPSS enthält die Partialkorrelationen, wenn auch mit negativem Vorzeichen)
- Bartlett-Test auf Sphärizität ist ein Maß für die Anwesenheit von signifikanten Interkorrelationen bei zumindest einigen Variablen
- MSA (measure of sampling adequacy) quantifiziert ebenfalls die Interkorrelationen
Kriterien zur Bewertung: MSA unter .50 ist unakzeptable, über .80 ist super
MSA steigt mit N, durchschnittlicher Interkorrelation, Anzahl der Variablen und abnehmender Anzahl der Faktoren

4. Schritt: Ableitung der Faktoren und Bestimmung des Fit

a) Allgemeine Faktorenanalyse (Hauptachsenanalyse) vs. Komponentenanalyse

- bei der Faktorenanalyse wird die Varianz in drei Teile aufgeteilt:
gemeinsame Varianz (allen Variablen gemeinsam)
spezifische Varianz (Varianz, die nur mit einer bestimmten Variable verbunden ist)
Fehlervarianz (Unreliabilität, Meßfehler, Zufallskomponente)
- die (Haupt-) Komponentenanalyse betrachtet die komplette Varianz; auf der Diagonalen in der Korrelationsmatrix werden Einsen eingetragen
- hingegen verwendet die „common factor analysis“ (Hauptachsenanalyse) nur die gemeinsame Varianz; auf der Diagonalen der Matrix werden Kommunalitäten eingetragen
Kommunalität: Varianz, die eine Variable mit allen anderen Variablen in der Analyse gemeinsam hat

Entscheidung für eins der beiden Modelle basiert auf folgenden Kriterien:

- Ziel der Faktorenanalyse
- Wissen über die Faktoren
- die Komponentenanalyse ist angemessen, wenn es um die Maximierung der Varianzaufklärung mit einer minimalen Anzahl an Faktoren geht, und wenn theoretische Überlegungen nahelegen, daß spezifische und Fehlervarianz sehr gering sind.
- die Hauptachsenanalyse ist hingegen geeignet, die latenten Dimensionen zu identifizieren, und wenn der Forscher wenig über die spezifische und Fehlervarianz weiß.

Probleme bei der Hauptachsenanalyse:

- manchmal gibt es keine eindeutige Lösung, das bedeutet, daß für eine Person verschiedene Faktorscores berechnet werden können
- Kommunalitäten sind nicht immer schätzbar
- die beiden Analysen bringen meist vergleichbare Ergebnisse, vor allem, wenn man mehr als 30 Variablen in die Analyse aufnimmt oder die Kommunalitäten der meisten Variablen größer als .60 sind
- der Gebrauch der Hauptkomponentenanalyse hat sich sehr durchgesetzt.

b) Kriterien für die Bestimmung der Faktorenanzahl

- generell sollte man stets mehrere Versionen ausprobieren; auch wenn man sich zum Beispiel vorher für drei Faktoren entschieden hat, sollte man die Lösungen mit zwei und vier Faktoren betrachten!
- Eigenwertkriterium: es sollten nur die Faktoren einbezogen werden, die einen Eigenwert >1 haben (das bedeutet, daß sie die Varianz von mehr als einer Variable erklären können); sehr reliabel bei 20 bis 50 Variablen
- A priori Kriterium: die Anzahl der Faktoren wird vorher aufgrund theoretischer Überlegungen festgelegt.

- Varianzanteilkriterium: die Faktoren sollten einen bestimmten Anteil der Varianz aufklären (z.B. 95% in den Naturwissenschaften; 60% in den Sozialwissenschaften)
- Scree test criterion: die Eigenwerte werden gegen die Anzahl der Faktoren geplottet; an der Stelle, an der die (abfallende) Funktion einen Knick macht, tragen die weiteren Faktoren mehr spezifische als gemeinsame Varianz bei und sollten nicht mehr einbezogen werden
- Heterogenität der Pbn: die ersten Faktoren sind stets die, die homogen über die ganze Stichprobe sind; möchte der Forscher Faktoren finden, die Subgruppen in der Stichprobe diskriminieren, sollte er zusätzliche Faktoren bestimmen

Folgen einer unangemessenen Faktorenanzahl:

- zu wenige Faktoren: die korrekte Struktur kann nicht aufgedeckt werden, möglicherweise werden wichtige Dimensionen übersehen
- zu viele Faktoren: Interpretation wird schwieriger, vor allem nach einer Rotation; Sparsamkeit ist in der Faktorenanalyse genau so von Bedeutung wie in anderen multivariaten Verfahren!

5. Schritt: Interpretation der Faktoren

- zunächst wird die unrotierte Ladungsmatrix betrachtet; ist der Forscher nur an der besten Linearkombination der Variablen interessiert, so kann er die hier ablesen
Faktorladung: Korrelation einer Variablen mit dem Faktor
- die unrotierte Lösung reduziert zwar die Daten, bietet jedoch nicht unbedingt ein bedeutsames Muster, das den Daten zugrunde liegt.
- die Faktoren können also rotiert werden, um eine bedeutsamere Struktur zu erhalten.

a) Rotation der Faktoren

- der Effekt der Rotation ist, die Varianz, die die einzelnen Faktoren erklären, umzuverteilen, bis ein bedeutsames Muster entsteht.
- orthogonale Rotation: die Faktoren bleiben unabhängig voneinander
- oblique Rotation: die Faktoren müssen nicht unabhängig voneinander bleiben (viel realistischer und flexibler!)

orthogonale Rotationen:

- hier werden entweder die Zeilen oder die Spalten (Faktoren) der Korrelationsmatrix vereinfacht (das bedeutet, daß die Werte entweder nahe Null oder möglichst hoch sein sollen).
- bei Vereinfachung der Zeilen soll die Ladung einer Variablen auf einem Faktor möglichst hoch werden und auf den anderen Faktoren möglichst gering
- Vereinfachung der Spalten bedeutet, daß auf jedem Faktor möglichst nur eine Variable hoch laden sollte und die anderen eher gering.
- Quartimax: die Zeilen sollen vereinfacht werden: jede Variable soll auf nur einem Faktor hoch laden; oft entsteht hierbei ein großer Faktor; nicht sehr erfolgreich
- Varimax: die Spalten sollen vereinfacht werden: die Varianz der Ladungen auf einem Faktor wird maximiert bzw. die Ladungen auf einem Faktor sind entweder sehr hoch oder sehr niedrig; Interpretation wird einfacher, es entsteht eine klare Trennung der Faktoren; sehr erfolgreiche Methode

- Equimax: Kompromiß von Quartimax und Varimax; versucht beides, wird aber kaum angewendet.

oblique (schiefwinklige) Rotationen:

- zum Beispiel Oblimin und PROMAX in SPSS

Die Entscheidung für die Form der Rotation sollte überlegt getroffen werden; meist nimmt der Forscher die Default-Variante Varimax.

b) Kriterien für die Signifikanz der Faktorladungen

- Faktorladungen sollten mindestens .30 betragen, ab .50 spricht man von praktischer Signifikanz
- die quadrierte Ladung ist der Anteil an der Varianz einer Variable, der durch den Faktor erklärt wird
- die statistische Signifikanz von Faktorladungen kann genau wie für Korrelationskoeffizienten bestimmt werden, jedoch haben Faktorladungen wesentlich größere Standardfehler; es gibt dazu konservative Richtlinien in Tabellen (wie groß muß mein N sein, damit bei bestimmter Power eine Ladung signifikant wird?)
- auch die Anzahl der Faktoren hat einen Einfluß auf die Beurteilung der Faktorladungen; es ist klar, daß bei zusätzlichen Faktoren die Ladungen keinen vergleichbar hohen Wert aufweisen müssen, um als bedeutsam bewertet zu werden!

c) Interpretation der Ladungsmatrix

- bei schiefwinkliger Rotation erhält man zwei Matrizen:
Faktorenmuster-Matrix: Ladungen, die den spezifischen Beitrag jeder Variable zu den Faktoren darstellen
Faktorenstruktur-Matrix: Korrelationen von Variablen und Faktoren, wobei die Interkorrelation der Faktoren mit eingehen
- zuerst sollte man die höchste Ladung für jeden Faktor markieren
- Variablen, die auf mehreren Faktoren hohe Ladungen aufweisen, erschweren die Interpretation und sind Kandidaten für eine Entfernung
- Betrachtung der Kommunalitäten: diese stellen den Anteil der Varianz einer Variablen dar, die durch den Faktor erklärt wird
- man kann z.B. festlegen, daß mindestens 50% der Varianz der einbezogenen Variablen erklärt werden sollen
- Variablen, die weder hohe Kommunalitäten noch hohe Ladungen aufweisen sollten entweder bei der Interpretation ignoriert oder ganz aus der Analyse entfernt werden!
- schließlich müssen für die Faktoren schöne Namen/Labels gefunden werden; bei orthogonalen Rotationen ist zu beachten, daß die Faktoren (also auch deren Interpretationen) unabhängig voneinander sind

6. Schritt: Validierung der Faktorenanalyse

- hier geht es um die Generalisierbarkeit der gefundenen Struktur
- z.B. mit Split Samples wird überprüft, inwiefern die Struktur replizierbar ist
- es gibt verschiedene Ansätze, mit denen man die Übereinstimmung von Ladungsmatrizen prüfen kann
- auch von Interesse ist die Stabilität der Faktoren (hängt hauptsächlich von der Stichprobengröße und den Beobachtungen pro Variable ab)
- einflußreiche Beobachtungen und Ausreißer sollten betrachtet werden (wie bei der Regression)

7. Schritt: Weitere Verwendungen der Ergebnisse einer Faktorenanalyse

- a) Auswahl einer „Surrogat“-Variablen für jeden Faktor
 - für die weitere statistische Verarbeitung
 - diese Variable soll repräsentativ für einen Faktor sein
 - hierfür wird meist die Variable mit der höchsten Ladung gewählt, wenn deren Reliabilität jedoch schlecht ist, kann auch eine reliablere Variable mit geringfügig geringerer Ladung gewählt werden
 - Nachteil: Informationsverlust oder verzerrte Ergebnisse bei weiteren Analysen
- b) Gestaltung einer „Summenskala“
 - aus mehreren Variablen (Indikatoren) wird gemäß den Ergebnissen der FA eine Skala gebildet; es wird die Linearkombination mehrerer Variablen zu einem Gesamtwert gebildet
 - der Meßfehler, der Einzelmessungen anhaftet, wird geringer
 - die verschiedenen Aspekte eines Konstrukt werden in einem Maß abgebildet
 - entspricht dem Postulat der Sparsamkeit
 - häufig in der Psychodiagnostik eingesetzt

 - wichtig ist die Inhalts- oder Augenscheinvalidität
 - Skala sollte eindimensional sein (ergibt sich aus der FA!)
 - Reliabilität der Skala: die Konsistenz der Skala kann mit Cronbach's Alpha bestimmt werden (sollte mindestens .60 sein!)
 - auch konvergente, diskriminante und nomologische (prognostische) Validität sollten abgesichert sein.
- c) Berechnung von Faktorscores
 - Faktorscore bildet ab, inwieweit eine Person auf den Items, die hoch auf dem Faktor laden, hohe Werte hat
 - im Gegensatz zur Summenskala wird der Faktorscore aus allen Variablen bzw. deren Ladungen bestimmt
 - Nachteil dieser Scores ist die schlechte Replizierbarkeit

Wenn die Summenskala gut konstruiert ist und den Gütekriterien entspricht, ist sie die beste Alternative!

Diskriminanzanalyse

1. Schritt: Inhalte der Diskriminanzanalyse

- Verständnis von Gruppenunterschieden oder Klassifikation von Objekten in Gruppen
- kategoriale abhängige und metrische unabhängige Variablen

2. Schritt: Untersuchungsdesign

- Auswahl der Variablen: zuerst sollte die abhängige (Gruppierungs-) Variable festgelegt werden; die Kategorien müssen erschöpfend und disjunkt sein
- beim Extrempol-Ansatz werden die Extremgruppen eines Kontinuums (metr. Variable) verglichen (nur geeignet, wenn bei der Regression nix gescheites rausgekommen ist)
- Stichprobengröße sollte mindestens 20 Fälle pro Prädiktor betragen; die DA reagiert sehr sensibel auf zu geringes N.
- die Gruppen sollten annähernd gleich groß sein (zu große Unterschiede beeinflussen die Schätzung)
- üblicherweise wird die Stichprobe in eine Analyse- und eine Holdout-Stichprobe geteilt (dabei sollte die Verteilung relevanter Merkmale gleich sein!)
- wenn N zu gering ist für diese Aufteilung, so kann die Validierung an der gleichen Stichprobe vorgenommen werden, was zu einer Überschätzung der Klassifikationsleistung führt

3. Schritt: Überprüfung der Annahmen

- multivariate Normalverteilung und ähnliche Streuungs- und Kovarianzstrukturen in den Gruppen
- wenn die multivariate NV nicht gegeben ist, ist die logistische Regression besser zur Analyse geeignet
- bei unterschiedlichen Kovarianzen werden die Beobachtungen häufiger in Gruppen mit größeren Kovarianzmatrizen geordnet; Abhilfe schafft hier die Vergrößerung der Stichprobe
- Multikollinearität verursacht vor allem bei schrittweisen Prozeduren Probleme
- Annahme linearer Beziehungen
- Ausreißer beeinflussen die Genauigkeit der Klassifizierung

4. Schritt: Schätzung der Diskriminanzfunktionen und Bestimmung des Modell-Fits

a) Berechnung

- es werden für n Gruppen n-1 Diskriminanzfunktionen geschätzt; für diese Funktionen wird jeweils ein Cut-off bestimmt, der über die Gruppenzuordnung entscheidet.
- man kann das Diskriminanzmodell auf verschiedene Arten schätzen:
- Simultane Schätzung: alle UVn werden auf einmal einbezogen

- Schrittweise Schätzung: zuerst wird die UV einbezogen, die die stärkste Diskriminierung ermöglicht; zu Beginn einbezogene Variablen können später auch wieder entfernt werden, wenn andere Variablen hinzukommen

b) Statistische Signifikanz

- Wilks Lambda, Hotellings Spur, Pillais Kriterium bewerten die Signifikanz der Diskriminationskraft der Diskriminationsfunktion(en)
- Roys „greatest characteristic root“ bezieht sich nur auf die erste Diskriminanzfunktion
- Mahalanobis D^2 und Raos V sind Maße für schrittweise geschätzte Modelle
- Mahalanobis-Distanz D^2 beruht auf der generalisierten euklidischen Distanz, adjustiert um ungleiche Varianzen; ermöglicht maximale Ausnutzung der vorhandenen Informationen; diese Distanz wird zwischen den Gruppen maximiert
- gibt es drei oder mehr Gruppen, so muß nicht nur die Signifikanz des ganzen Modells geprüft werden, sondern auch die Bedeutsamkeit jeder einzelnen Funktion
- ist eine Funktion nicht signifikant, so sollte das Modell erneut geschätzt werden, wobei die Anzahl der Funktionen begrenzt werden sollte auf die Anzahl signifikanter Funktionen im ersten Modell

c) Bestimmung des Modell-Fit

Berechnung diskriminanter Z-Werte

- für jede Person kann pro Diskriminationsfunktion ein Z-Wert berechnet werden, der einen Vergleich mit anderen Personen ermöglicht.
- auf Basis dieser Z-Werte werden die Personen in bestimmte Gruppen geordnet
- im Gegensatz zur Diskriminanzfunktion ergibt die Klassifikationsfunktion (auch als Fishers lineare Diskriminanzfunktion) für jede Person einen Klassifikationswert für jede Gruppe. Die Person wird dann in die Gruppe geordnet, für die sich der höchste Klassifikationswert ergibt.

Bewertung von Gruppenunterschieden

- die Gruppenzentroiden sind im Prinzip die Mittelwerte der einzelnen Gruppen
- Gruppenzentroiden sind der durchschnittliche Z-Wert aller Gruppenmitglieder
- zur Bewertung von Gruppenunterschieden kann die Mahalanobis-Distanz der Zentroiden bestimmt werden
- die Lage der Gruppen kann auch in einer territorialen Karte dargestellt werden (immer für zwei Diskriminanzfunktionen)

Beurteilung der Genauigkeit der Klassifikationsleistung

- es werden Klassifikationstabellen aufgestellt, die die Trefferquote bei der Klassifikation der Personen erkennen lassen
- um die Klassifikationstabelle aufstellen zu können, müssen die Cut-offs für die einzelnen Diskriminanzfunktionen bestimmt werden.
- dieser optimale (kritische) Z-Wert liegt bei gleich großen Gruppen genau in der Mitte zwischen den Gruppenzentroiden; sind die Gruppen ungleich groß, muß ein gewichteter Mittelwert bestimmt werden.
- Als kritischer Wert wird der gewählt, der zum geringsten Anteil an Mißklassifikationen führt.
- die Klassifikationsmatrix beruht auf der Klassifikation in der Hold-Out-Stichprobe

- zur Vorhersagegüte kann ein t-Test berechnet werden:

$$t = \frac{p - 0.5}{\sqrt{\frac{0.5(1-0.5)}{N}}}, \text{ wobei } p \text{ der Anteil korrekter Vorhersagen ist.}$$

Vergleich der Vorhersagen mit den zufällig erreichten korrekten Vorhersagen:

- bei gleich großen Gruppen werden per Zufall in jede Gruppe gleich viele Personen geordnet
- Maximum chance criterion: wenn der Anteil korrekter Klassifikationen nicht den Anteil übersteigt, den die größte Gruppe an der ganzen Stichprobe hat, ist die Diskriminanzanalyse nicht erfolgreich; angemessenes Kriterium, wenn es darum geht, den Anteil korrekter Vorhersagen zu maximieren.
- Proportional chance criterion: wenn bei ungleich großen Gruppen die Zuordnung nicht nur zur größten Gruppe möglichst gut sein soll;
 $C_{\text{PRO}} = p^2 + (1-p)^2$, wobei p der Anteil einer Gruppe ist.
- Kriterium zur Güte der Klassifikation: die Vorhersagen auf Basis der Diskriminanzfunktionen sollten mindestens ein Viertel besser sein als bei Zufall erwartet.
- Press' Q-Statistik: statistischer Test zur Überprüfung der Diskriminanzleistung im Vergleich zu einem Zufallsmodell;

$$Q = \frac{[N - (nK)]^2}{N(K-1)}, \text{ wobei } K \text{ die Anzahl der Gruppen ist und } n \text{ die Anzahl korrekt klassifizierter Personen; der kritische Wert wird einer Chi-Quadrat-Verteilung mit } df=1 \text{ entnommen.}$$

d) Einzelfalldiagnostik

- Fehlklassifikation einzelner Personen
- Versuch, Charakteristika dieser Personen zu erkennen, die vielleicht systematisch zu Fehlklassifikationen geführt haben; diese Variablen könnten evtl. mit aufgenommen werden als UVn
- nichtlineare Zusammenhänge könnten auffallen
- je weiter ein Fall vom Gruppenzentrum entfernt liegt, desto weniger ist er repräsentativ für die Gruppe
- die fehlerklassifizierten Fälle könnten auch in einer territorialen Karte betrachtet werden

5. Schritt: Interpretation der Ergebnisse

- die Diskriminanzgewichte (auch -koeffizienten) können wie die Beta-Gewichte in der Regression als Beitrag der einzelnen Variablen zur jeweiligen Diskriminanzfunktion interpretiert werden
- die Diskriminanzladungen (auch Strukturkoeffizienten genannt) sind die einfache Korrelation der UVn mit den Faktoren; sie können entsprechend den Faktorladungen in der Faktorenanalyse interpretiert werden.
- es können partielle F-Werte bestimmt werden, was vor allem bei schrittweisen Prozeduren sinnvoll ist.
- die Diskriminanzfunktionen können wie bei der Faktorenanalyse rotiert werden (meist Varimax), um die Varianz umzuverteilen

- es kann ein Potenzindex für jede Variable bestimmt werden, der ein Maß für ihre diskriminierende Kraft über alle Diskriminanzfunktionen hinweg darstellt; das ist jedoch nur ein relatives Maß!
Es wird erst ein Potenzindex für jede Funktion bestimmt:
 $\text{Potenz} = (\text{Diskriminanzladung})^2 \times \text{relativer Eigenwert der Funktion}$
Diese Potenzwerte können dann für alle Funktionen summiert werden.
- die Möglichkeit einer graphischen Darstellung besteht auch: „stretching the vectors“

Wahl der geeigneten Interpretationsmethode: die Interpretation von Diskriminanzladungen ist am zuverlässigsten!

6. Schritt: Validierung der Ergebnisse

- Kreuzvalidierung, wobei der Gebrauch einer Hold-Out-Stichprobe die Generalisierbarkeit erhöht!
- alternativ kann auch Jackknifing betrieben werden
- das Erstellen von Gruppenprofilen ermöglicht ein Urteil darüber, ob die Ergebnisse mit den konzeptuellen oder theoretischen Überlegungen konsistent sind.

Logistische Regression

- robustes Verfahren, das bei Verletzungen der NV-Annahme nicht so sensibel reagiert
- für binäre abhängige Variablen
- die Ähnlichkeit zur Regression schließt zahlreiche statistische Tests mit ein, die gerechnet werden können
- es wird direkt eine Wahrscheinlichkeit vorhergesagt; auch wenn eine WK eine metrische Variable ist, wird im Gegensatz zur multiplen Regression eine nichtlineare, nämlich die logistische Verteilung zugrundegelegt.
- Fehlerterme sind binomialverteilt
- Heteroskedastizität wird erwartet
- die logistische Funktion wird in einer Maximum-Likelihood-Schätzung den Daten angepaßt
- vgl. Logit-Analyse:
 $\ln(p/1-p) = b_0 + b_1x_1 + \dots + b_nx_n$

Bestimmung des Modell-Fits:

- die Güte eines Modells wird über die LogLikelihood bestimmt, die möglichst gering sein sollte
- verschiedene Modelle können mittels Chi-Quadrat verglichen werden, woraus sich auch ein „Pseudo- R^2 “ ableiten läßt:
$$R_{\logit}^2 = \frac{-2\text{LogLikelihood}_{null} - (-2LL_{\text{model}})}{-2\text{LogLikelihood}_{null}}$$
- Testung der Signifikanz der Koeffizienten: Wald Statistik (Test für die Signifikanz jedes einzelnen Koeffizienten)