Faktorenanalyse

Begriffe

Faktoren
: latente, d.h. nicht-meßbare Variablen

Indikatoren
: manifeste, d.h. meßbare, gemessene Variablen

Konstrukte
: latente Variablen, die nicht auf Indikatoren wirken

exogene Variablen
: Variablen, die nicht durch andere Variablen des Modells beeinflußt werden, die nur auf andere Variablen wirken.

endogene Variablen
: Variablen, die durch andere Variablen des Modells beeinflußt werden, die aber auch auf andere Variablen wirken können.

konfirmatorische Faktorenanalyse
: ein Modell der Beziehungen zwischen Faktoren und Indikatoren wird (qualitativ und/oder quantitativ) vorgegeben, die Modellparameter bestimmt und eventuelle Anpassungstests durchgeführt; insbesondere liegen Anzahl der Faktoren und Zuordnung der Indikatoren zu den Faktoren von vorneherein fest.

exploratorische Faktorenanalyse
: es wird kein Modell der Beziehungen und i.a. keine Faktorenzahl vorgegeben, jeder Indikator kann mit jedem Faktor in Beziehung stehen; alle möglichen Beziehungen werden quantitativ ermittelt; Anpassungstests sind i.a. nicht möglich.

Mathematische Symbole und Beziehungen

n	Anzahl der Objekte, Versuchspersonen
m	Anzahl der Variablen, Indikatoren
(z_{ij})	eine Datenmatrix (mit i=1,2,...,n und j=1,2,...,m) metrischer Variablen, die als standardisiert angenommen werden, d.h. $\bar{x}(z_j) = 0$ und $s(z_j) = 1$
z_j	eine Variable
(r_{jk})	die Korrelationsmatrix der m Variablen (mit j=1,2,...,m und k=1,2,...,m)
f_p	eine latente Variable, Faktor oder Konstrukt (p=1,2,...,r)
(f_{ip})	die Datenmatrix der Faktoren (mit i=1,2,,,.n und p=1,2,...,r), die sog. *Faktorscore-Matrix*

Faktorladungen:
: a_{jp} quantitativer Einfluß des Faktors auf den Indikator,
in der Regel Korrelation der Variablen j mit dem gemeinsamen Faktor p ($r_{z_j f_p}$)

Modell der Hauptkomponenten-Analyse (PCA, principal component analysis)
: $z_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jr}f_r$ für j=1,2,...,m

Modell der Hauptfaktoren-Analyse (PFA, principal factor analysis)

$z_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jr}f_r + d_j u_j$ für j=1,2,...,m

mit einem *spezifischen Faktor* u_j

Streuungszerlegung

$s_j^2 = 1 = \sum_1^n \frac{z_{ij}^2}{n} = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jr}^2 + [d_j^2]$

Kommunalität

$h_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jr}^2$

Anteil der durch die r Faktoren erklärte Varianz der Variablen j

insbesondere gilt: $h_j^2 = 1$ bei PCA und $h_j^2 < 1$ bei PFA

Einzelvarianz:

$d_j^2 = 1 - h_j^2$

Anteil der durch die r gemeinsamen Faktoren nicht erklärte Varianz der Variablen j

Faktorvarianz:

$V_p = a_{1p}^2 + a_{2p}^2 + \ldots + a_{mp}^2$

Anteil der Gesamtvarianz m, die der Faktor p erklärt

$V = V_1 + V_2 + \ldots + V_r$

Anteil der Gesamtvarianz m, die durch die r gemeinsamen Faktoren erklärt wird

Faktorbeziehungen:

$\bar{f_i} = 0$ $s_{f_i} = 1$ standardisiert (Mittelwert = 0, Varianz = 1)

$r_{f_i f_j} = 0$ $i \neq j$ orthogonal / unkorreliert

Faustregeln zur Bestimmung der Faktorenzahl

- Ein Varianzanteil (bezogen auf die Gesamtvarianz m) soll erklärt werden; allerdings nur sinnvoll bei PCA, da hier keine spezifischen Faktoren mit Einzelvarianz auftreten, die durch die gemeinsamen Faktoren nie erklärt werden können.
  (z.B. 90% der Gesamtvarianz)
- Ein Varianzanteil bezogen auf die Gesamtkommunalität ($h_1^2 + h_2^2 + \ldots + h_m^2$) soll erklärt werden; sinnvoll bei PFA (bei PCA wegen h_j=1 identisch mit o.a. Kriterium).
- Ein Varianzanteil eines Faktors ($V_p = a_{1p}^2 + a_{2p}^2 + \ldots + a_{mp}^2$) sollte größer als 1 sein (1 ist der Durchschnittswert von V_p bei r=m), d.h. es werden so viele Faktoren gebildet wie Eigenwerte $\lambda > 1$ vorhanden sind (*Kaiser-Kriterium*). Bei kleiner Variablenzahl werden zu wenige, bei großer Variablenzahl zu viele Faktoren gbeildet.
- Scree-Test: Die Eigenwerte werden der Größe nach angeordnet. Bei unabhängigen Zufallsvariablen (d.h. wenn keine Faktorstruktur vorhanden ist) liegen diese auf einer Geraden um 1. Danach Einzeichnen einer Geraden (von rechts) durch die letzten Eigenwerte. Die Eigenwerte (links), die deutlich oberhalb der Geraden liegen, entsprechen echten Faktoren.
- Standard-Strategie:
  r (Anzahl der Faktoren) < m/2
  nur Faktoren p mit V_p>1
  $V_1 + V_2 + \ldots + V_r \geq 0,9(h_1^2 + h_2^2 + \ldots + h_m^2)$,d.h.
  Faktoren erklären mindestens 90% der durch die gemeinsamen Faktoren erklärbare Varianz.

Rotation

Die Faktorenlösung ist nicht eindeutig. Aus den unendlich vielen wird durch Drehung der Faktorenachsen (Rotation), also einer Änderung der Orientierung der Faktoren, eine „gut erklärbare“ ausgesucht.

- Varimax-Rotation:
  Durch die Maximierung der Varianz der Faktorladungen a_{jp} wird erreicht erreicht, dass möglichst viele Ladungen entweder Werte nahe bei 1 oder -1 bzw. bei 0 haben, also die Faktorachsen möglichst durch die Variablenpunkte verlaufen. Dadurch sind Faktoren und Variablen gut zuzuordnen.
- Oblimin-Rotation:
  Hierbei wird die Orthogonalität der Faktoren, die noch bei der Faktorextraktion angenommen wurde, aufgehoben, d.h. die Faktoren dürfen miteinander korrelieren. Sollte nur angewandt werden, wenn ein Zusammenhang der Faktoren erklärbar ist.

Durch die Rotation verschieben sich die Varianzanteile der Faktoren. Da die Rotation nach der Faktor-Extraktion und der Festlegung der Faktorenzahl durchgeführt wird, bleibt die durch die Faktoren erklärte Varianz konstant. Allerdings werden die Anteile gleichmäßiger auf die Faktoren aufgeteilt.

Voraussetzungen

Die Korrelationsmatrix sollte dem Skalenniveau entsprechen, z.B.

- Produkt-Moment-Korrelationen bei ausschließlich metrischen (und einigen dichotomen) Merkmalen
- Spearman-Rangkorrelationen bei Auftreten von ordinalen Merkmalen
  (in SPSS: Rangfolgen bilden vor Berechnung von Produkt-Moment-Korrelationen)
- Nominale Merkmale müssen dichotomisiert werden
  (k Ausprägungen ergeben (k-1) dichotome Variablen.)
- Phi-Koeffizienten bei ausschließlich dichotomen Merkmalen
  (in SPSS: simple Berechnung von Produkt-Moment-Korrelationen)

Korrelations/Kovarianzmatrix muss *positiv-semidefinit* sein. Dies ist i.a. gegeben wenn

- Fallzahl (n) > Variablenzahl (m) + 2
  Faustregel: n >> 2*m
- keine fehlenden Werte
- Variablen dürfen nicht *multicollinear* sein.
  (Variablen x_0, x_1, x_2, x_3,.. sind multicollinear, wenn sich eine der Variablen, z.B. x_0, annähernd als Linearkombination der übrigen Variablen darstellen lässt:
  $x_0 = a_0 + a_1 x_1 + a_2 x_2 + \ldots$
  (in SPSS überprüfbar über eine lineare Regression)