

Georg Quaas (Dezember 2008 / Januar 2013)

Faktorenanalyse (Principal Components Analysis) : Mathematischer Hintergrund

(Dazugehörige Beispielrechnung: File BspAbi.xls)

Die Matrix der p beobachteten Variablen x_1, x_2, \dots, x_p über n Fälle ist:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{np} \end{pmatrix}$$

Sie soll in eine Matrix der Faktorwerte über $q < p$ Faktoren f_1, f_2, \dots, f_q und n Fälle:

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdot & \cdot & \cdot & f_{1q} \\ f_{21} & f_{22} & \cdot & \cdot & \cdot & f_{2q} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ f_{n1} & f_{n2} & \cdot & \cdot & \cdot & f_{nq} \end{pmatrix}$$

überführt werden. Die Informationsreduktion drückt sich in der Ungleichung $q < p$ aus. Die folgende Matrix der Faktorladungen enthält die Koeffizienten für die Linearkombinationen, die die Faktorwerte in beobachtete Werte überführen:

$$A' = \begin{pmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1p} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{q1} & a_{q2} & \cdot & \cdot & \cdot & a_{qp} \end{pmatrix}$$

so dass gilt:

$$X = F \cdot A'$$

oder in Summenschreibweise

$$x_{mi} = \sum_{j=1}^q f_{mj} a_{ij}$$

Legende des allgemeinen Ansatzes:

:

m = Laufindex der Fälle: $1, 2, \dots, n$

i = Laufindex der p Indikatoren (der ursprünglichen Variablen)

j = Laufindex der q Faktoren

x_{mi} = Wert des m -ten Falles auf dem i -ten Indikator (Rohdatenwert)

f_{mj} = Wert des m -ten Falles auf dem j -ten Faktor (Faktorwert)

a_{ij} = Bedeutung des j -ten Faktors für den i -ten Indikator

Für Anwender von LISREL (Konfirmative Faktoranalyse) ist zu beachten, dass in den entsprechenden Handbüchern die Dimension der Fälle nicht dargestellt wird. Der folgende Abschnitt beschäftigt sich kurz mit den Besonderheiten der LISREL-Notation.

Der LISREL-Ansatz für ein einstufiges Faktorenmodell

Dem LISREL-HB, S.139, entsprechend gilt (ohne Fehlerterm):

$$x = \Lambda \xi$$

Achtung! In der LISREL-Notation fällt die Numerierung der Fälle weg:

Legende/Parallelität:

$$x' = [x_{(m)i}]$$

$$\Lambda = A$$

$$\xi' = [f_{(m)j}]$$

Alternative Notation:

$$x = A f' \text{ mit}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad f' = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_q \end{bmatrix}$$

Anmerkung:

Die a_{ij} sind die Korrelationen zwischen x_i und f_j .

Anmerkung:

Der allgemeine Ansatz für eine explorative Faktorenanalyse kann pfadanalytisch wie folgt charakterisiert werden: Von jedem Faktor geht ein Pfad a_{ij} zu jedem Indikator.

Die Begriffe der Kommunalität und des Eigenwertes lassen sich anhand eines LISREL-Messmodells (Beispiel S.135) wie folgt erläutern: Sei

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ 0 & \lambda_{32} \\ 0 & \lambda_{42} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix}$$

Anmerkung:

Ein LISREL-Messmodell unterscheidet sich vom Modell der explorativen Faktorenanalyse u.a. dadurch, dass einige Pfade = Null gesetzt werden (a priori-Spezifikation des Modells).

Vorausgesetzt, alle Variablen sind z-standardisiert, stellt die Summe der Quadrate der Elemente einer Zeile die Summe der quadrierten Ladungen einer Variablen auf allen Faktoren dar, d.h. die Kommunalität:

$$0 \leq h_i^2 = \sum_{j=1}^q \lambda_{ij}^2 = \sum_{j=1}^q a_{ij}^2 \leq 1$$

Die Kommunalität gibt an, in welchem Maße die Varianz einer beobachteten Variable durch die Faktoren aufgeklärt wird.

Dagegen ergibt die entsprechende Spaltensumme den Eigenwert eines Faktors, das heißt die Gesamtvarianz aller Variablen, die durch diesen Faktor aufgeklärt wird (höchstens p):

$$\lambda_j = \sum_{i=1}^p \lambda_{ij}^2 \leq p.$$

Ende Exkurs zur LISREL-Notation

Der folgende Text stützt sich u.a. auf Bortz, insbesondere auf die Formel 15.3b:

$$X = F A'$$

An die Stelle des Vektors x der LISREL-Notation tritt hier wieder die Datenmatrix X , deren Zeilen die einzelnen Fälle darstellen. Die Bedingung der Dimensionsreduktion $q < p$ wird bei der folgenden Darstellung der Hauptschritte des Rechenganges nicht benutzt; vielmehr muss $q = p$ unterstellt werden, um diese „Hauptschritte“ zu verstehen.

Vor der Extraktion der Faktoren, die mathematisch betrachtet eine Rotation im p -dimensionalen Raum und anschließender Dimensions-Reduktion ist, werden die Werte auf

jeder Achse z-standardisiert. Die Faktoren werden so konstruiert, dass sie sukzessive „maximale Varianz“ aufklären und die Faktorwerte untereinander nicht korrelieren. In der graphischen Darstellung der Faktorwerte ist folglich eine Punktwolke zu sehen, die in jeder Richtung die gleiche Varianz hat. Der Vorteil der Standardisierung besteht darin, dass am Faktorwert unmittelbar abgelesen werden kann, „wie stark die in einem Faktor zusammengefassten Merkmale bei dieser Vp [Versuchsperson] ausgeprägt sind.“ (S.503)

Für das Verständnis der Hauptschritte einer Faktorenanalyse sind die folgenden drei Nebenbetrachtungen hilfreich:

(1) Mit Hilfe der Datenmatrix X bilden wir:

$$D = X'X - \bar{X}'\bar{X}$$

wobei

$$\bar{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \dots & \dots & \dots & \dots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix}$$

$$\text{Sei } e_n = e = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$$

ein Summationsvektor, so gilt:

$$\frac{1}{n}e'X = [\bar{x}_1 \quad \bar{x}_2 \quad \dots \quad \bar{x}_p]$$

$$\text{sowie } \frac{1}{n}ee'X = \bar{X}$$

und folglich

$$\bar{X}'\bar{X} = X'ee'\frac{1}{n^2}ee'X = X'e\frac{1}{n}e'X = X'\bar{X} = \bar{X}'X \quad (1)$$

(2) **Definition der Kovarianz- und der Korrelationsmatrix:**

$$\text{COV} = \frac{1}{n}D,$$

$$S = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & s_p \end{bmatrix},$$

$$R = S^{-1} \cdot \text{COV} \cdot S^{-1}$$

(3) z-Standardisierung der Messwerte:

$$Z = (X - \bar{X}) S^{-1} \text{ (Definition)}$$

Es gilt:

$$\begin{aligned} \frac{1}{n} Z'Z &= \frac{1}{n} S^{-1} (X' - \bar{X}') (X - \bar{X}) S^{-1} = \frac{1}{n} S^{-1} (X'X - \bar{X}'X - X'\bar{X} + \bar{X}'\bar{X}) S^{-1} \\ &= \frac{1}{n} S^{-1} (X'X - \bar{X}'\bar{X} - \bar{X}'\bar{X} + \bar{X}'\bar{X}) S^{-1} = \frac{1}{n} S^{-1} (X'X - \bar{X}'\bar{X}) S^{-1} \quad \text{wegen (1)} \end{aligned}$$

$$= \frac{1}{n} S^{-1} D S^{-1} = S^{-1} \text{COV} S^{-1} = R$$

$$\rightarrow R = \frac{1}{n} Z'Z \quad (2)$$

Die Hauptschritte einer PCA-Faktorenanalyse

An die Stelle der Rohdaten X und der Formel 15.3b (Bortz) treten die standardisierten Daten Z und damit die Formel

$$Z = F A', \quad Z, F \text{ sind standardisiert.} \quad (3)$$

Die Standardisierung hat den Vorteil, dass die aufzuklärende Varianz jeder Variable z gleich Eins ist.

In standardisierten Werten:

$$Z = F A'$$

$$n \text{ Zeilen: } \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ 1 & * & p \end{bmatrix} = \begin{bmatrix} * & * \\ * & * \\ * & * \\ * & * \\ * & * \\ 1 & q \end{bmatrix} \begin{bmatrix} * & * & * \\ * & * & p \end{bmatrix} q \text{ Zeilen}$$

Ziel einer Faktorenanalyse ist es, das Koordinatensystem, in dem die standardisierten Daten Z dargestellt werden, so zu drehen, dass durch das neue, gedrehte Koordinatensystem sukzessive maximale Varianz aufgeklärt wird. Aus der Matrixalgebra (Gantmacher, S.259) ist bekannt, dass eine Drehung im euklidischen Raum mit einer orthogonalen Matrix V realisiert wird, die die Eigenschaft hat, die Metrik des Raumes zu erhalten:

$$\langle Vz, Vz \rangle = \langle V'Vz, z \rangle = \langle z, z \rangle \Leftrightarrow V'V = E$$

Aus der Matrixgleichung $V'V = E$ folgt $|V'| |V| = 1 \Leftrightarrow |V| = \pm 1$. Der Fall einer Spiegelung $|V| = -1$ wird ausgeschlossen. Der erste (linke) Spaltenvektor v der Matrix V stellt die Koordinaten der ersten Koordinatenachse (Faktor 1) des gedrehten Koordinatensystems im alten Koordinatensystem dar. Die Varianz des Vektors

$$y = Zv,$$

der die Projektion der Daten Z auf die neue 1. Achse (Faktor 1) darstellt, soll maximal sein, also

$$\frac{1}{n} y'y = \frac{1}{n} v'Z'Zv = v'Rv = \lambda v'v.$$

Maximierung der Kovarianz unter der Nebenbedingung $v'v = 1$ mit Hilfe eines Lagrange-Multiplikators bedeutet:

$$\frac{\partial F}{\partial v} = \frac{\partial [v'Rv - \lambda(v'v - 1)]}{\partial v} = 2Rv - 2\lambda v = 0,$$

führt also auf das bekannte Problem der Bestimmung der Eigenvektoren v und der Eigenwerte λ der Gleichung

$$Rv = \lambda v.$$

Analysiert wird die charakteristische Gleichung der Korrelationsmatrix R :

$$|R - \lambda E| = 0 \rightarrow p \text{ Eigenwerte } \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

Zu jedem Eigenwert gibt es einen Eigenvektor v_i :

$$(R - \lambda_i E)v_i = 0, \rightarrow v_1, v_2, \dots, v_p,$$

die entsprechend der Größe der Eigenwerte (charakteristische Wurzeln) geordnet werden. Darüber hinaus werden alle Eigenvektoren orthogonalisiert, normiert und zur Matrix V zusammengefasst:

$$V = \begin{bmatrix} v_1 & v_2 & \dots & v_p \end{bmatrix}$$

Für die Rotationsmatrix V gilt:

$$RV = V\Lambda, \text{ wobei } \Lambda = \|\delta_{ik} \lambda_k\| \text{ und} \quad (4)$$

$$V'V = E \text{ und } |V| = 1 \text{ (Bed. für die Konstruktion von V).} \quad (5)$$

Die Transformation der z-Variablen in das neue Koordinatensystem erfolgt nun in 2 Schritten:

$$(i) \quad Y = ZV \quad (\text{Rotation}) \quad (6)$$

$$(ii) \quad F = Y\Lambda^{-1/2} \quad (\text{Normierung}) \quad (7)$$

Abgeleitete Eigenschaften:

Folge (i):

Die Faktorwerte F sind standardisiert und korrelieren untereinander nicht.

Beweis:

$$\frac{1}{n} F'F = \frac{1}{n} \Lambda^{-1/2} Y'Y \Lambda^{-1/2} \quad \text{wegen (7)}$$

$$= \frac{1}{n} \Lambda^{-1/2} V'Z'ZV \Lambda^{-1/2} \quad \text{wegen (6)}$$

$$= \Lambda^{-1/2} V'RV \Lambda^{-1/2} \quad \text{wegen (2)}$$

$$= \Lambda^{-1/2} V'V \Lambda \Lambda^{-1/2} \quad \text{wegen (4)}$$

$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = E \quad \text{wegen (5).}$$

Folge (ii):

Die Matrix A' enthält die Korrelationen zwischen Z und F , genannt „Faktorladungen“.

Beweis:

$$R_{ZF} = \frac{1}{n} F' Z = \frac{1}{n} F' F A' = A'$$

„Aus der Elementarstatistik wissen wir, dass das Quadrat einer Korrelation den Anteil gemeinsamer Varianz zwischen den korrelierten Meßwertreihen angibt. Das Quadrat der Ladung (a_{ij}^2) einer Variablen i auf einem Faktor j kennzeichnet somit den gemeinsamen Varianzanteil zwischen der Variablen i und dem Faktor j . Summieren wir die quadrierten Ladungen einer Variablen i über alle Faktoren, erhalten wir einen Wert h^2 , der angibt, welcher Anteil der Varianz einer Variablen durch die Faktoren aufgeklärt wird.“ (Bortz, S.504)

Demnach ist die Summe über die Quadrate der Elemente einer Zeile von A

$$0 \leq h_i^2 = \sum_{j=1}^q a_{ij}^2 \leq 1.$$

h_i^2 heisst die „**Kommunalität** der Variablen i “, und sie gibt an, in welchem Maße die Varianz dieser Variablen durch die Faktoren aufgeklärt bzw. erfasst wird.

Folge (iii):

Berechnung von A' , ohne dass die Faktorwerte benutzt werden:

Aus (6) und (7) folgt: $F \Lambda^{1/2} = Y = ZV$, also

$$F \Lambda^{1/2} V^{-1} = Z = F A' \quad \text{wegen (3)}$$

$$\rightarrow A' = \Lambda^{1/2} V^{-1} = \Lambda^{1/2} V'$$

Folge (iv):

Zusammenhang zwischen Faktorladungen und Matrix der Eigenwerte:

$$A' A = \Lambda^{1/2} V' V \Lambda^{1/2} = \Lambda \quad \text{wegen (5)}$$

Demnach gilt für die Summe der Quadrate einer Spalte von A :

$$\sum_{i=1}^p a_{ij}^2 = \lambda_j$$

Bedenkt man, dass das Quadrat einer Korrelation den Anteil der gemeinsamen Varianz zwischen den korrelierten Daten angibt, im vorliegenden Fall also zwischen den standardisierten Werten Z und den Faktorwerten F , so ergibt sich folgender Satz:

Die Summe der durch den j -Faktor aufgeklärten Gesamtvarianz aller Variablen ist $= \lambda_j$.

Folge (v):

Berechnung der Faktorwerte F :

$$F = ZV\Lambda^{-1/2} = ZV\Lambda^{1/2}\Lambda^{-1} = ZA\Lambda^{-1}$$

Folge (vi):

Zusammenhang zwischen Faktorladungen und Korrelationsmatrix:

$$\frac{1}{n}Z'Z = \frac{1}{n}AF'FA' = AA' = R$$

Die Hauptschritte des bis hier dargestellten „Rechenganges“ der Faktorenanalyse haben von der Bedingung $q < p$ (Informationsreduktion) keinen Gebrauch gemacht. Bei der PCA geht man davon aus, dass Faktoren, deren Eigenwert (charakteristische Wurzel) kleiner als Eins ist, keinen wesentlichen Beitrag zur Aufklärung der Gesamtvarianz mehr leisten und deshalb weggelassen werden können. Darin besteht die Informationsreduktion, auf die es bei der PCA ankommt.

Literatur:

Jürgen Bortz: Statistik für Sozialwissenschaftler. S.495 ff.

LISREL-Handbuch

F.R.Gantmacher: Matrizenrechnung. Band 1.