

## 9 Faktorenanalyse

Ziel der Faktorenanalyse ist es, die Anzahl der Variablen auf wenige voneinander unabhängige Faktoren zu reduzieren und dabei möglichst viel an Information zu erhalten. Hier wird davon ausgegangen, dass die Ausgangsvariablen mit latenten Variablen, den so genannten Faktoren, korrelieren. Bestimmte Variablen werden stärker mit bestimmten Faktoren korrelieren als andere. Handelt es sich beispielsweise um Daten aus der Psychologie oder den Sozialwissenschaften, so können die entsprechenden Ausgangsvariablen zu Gruppen zusammengefasst werden. Die Interpretation der entsprechenden Faktoren ist eine Angelegenheit des Psychologen bzw. Sozialwissenschaftlers. Bei einer Untersuchung (zum Beispiel im Rahmen eines psychologischen Tests) könnten die einzelnen Faktoren als bestimmte latente Persönlichkeitseigenschaften interpretiert werden.

Eine Faktorenanalyse ist auch sinnvoll, falls im Rahmen einer Regressionsanalyse der Einfluss mehrerer unabhängiger Variablen auf eine abhängige Variable untersucht werden soll. Oft kommt es vor, dass die unabhängigen Variablen in der Modellgleichung untereinander korrelieren. Hier könnte zunächst eine Faktorenanalyse mit diesen Variablen durchgeführt werden. Die Faktoren, die dabei extrahiert werden, könnten dann als neue unabhängige Variablen im Regressionsmodell verwendet werden.

Wir gehen zunächst von dem folgenden Modell aus (Modell der Hauptkomponentenanalyse):

$$Z = F L^T$$

F ist die unbekannte Faktorenmatrix (F ist eine orthogonale Matrix) und L die unbekannte Ladungsmatrix und  $L^T$  deren Transponierte. Unser Modell ähnelt dem linearen Modell der Regressions- bzw.

Varianzanalyse. Der Unterschied zum linearen Modell der Regressions- bzw. Varianzanalyse besteht darin, dass es zunächst keine Fehlermatrix  $E$  gibt und dass beide Matrizen auf der rechten Seite der Modellgleichung unbekannt sind. Die Matrix  $F$  entspricht dabei der Designmatrix  $X$ , und die Matrix  $L^T$  entspricht der unbekannt Parametermatrix  $\beta$ . Die Matrix  $Z$  ergibt sich aus der Datenmatrix  $Y$ , indem die Spalten von  $Y$  standardisiert werden (auf eine spezielle Art, wie wir gleich sehen werden). Da wir von einem orthogonalen Faktorenmodell ausgehen, ergibt das Produkt der Matrizen  $F^T F$  die Einheitsmatrix mit  $k$  Zeilen und Spalten.

Im ersten Schritt bestimmen wir die beiden unbekannt Matrizen  $F$  und  $L$  vollständig, was in der Literatur als Hauptkomponentenanalyse bezeichnet wird. Im zweiten Schritt kommen wir zur eigentlichen Faktorenanalyse, wobei wir die Anzahl der Faktoren  $k$  reduzieren, die im ersten Schritt noch mit der Variablenanzahl  $p$  identisch ist, so dass wir mit diesen extrahierten Faktoren möglichst viel der Varianz der Datenmatrix  $Y$  bzw.  $Z$  erklären. Falls die Zahl der Faktoren (= Spaltenanzahl der Matrix  $F$ ) nicht reduziert wird, erklärt das Modell die Daten vollständig (analog zur Regressionsanalyse  $Y = X\beta$  mit  $E = 0$ ). Werden  $k < p$  Faktoren extrahiert (i. A. über das Kaiserkriterium, das wir später erklären), so muss auf der rechten Seite der Modellgleichung noch die Fehlermatrix  $E$  hinzu addiert werden. Wir machen hierbei keine weiteren Verteilungsannahmen wie in der Regressions- und Varianzanalyse (es können aber auch Annahmen im Rahmen der Faktorenanalyse gemacht werden).

Die Faktorenanalyse ist immer dann sinnvoll, wenn mehrere Variablen erfasst wurden, die untereinander abhängig sind.

Wir werden im Folgenden an einem Beispiel die Faktorenanalyse durchführen. Danach stellen wir noch einen Test vor, mit dem überprüft werden kann, ob die Ausgangsvariablen signifikant korrelieren. Dieser sollte in der Praxis zu Beginn einer Faktorenanalyse oder Hauptkomponentenanalyse durchgeführt werden.

In unserem Beispiel gehen wir davon aus, dass bei 5 Schülern die Benotung in den Fächern Mathematik (erste Spalte der Datenmatrix **Y**), Physik (zweite Spalte) und Biologie (dritte Spalte) in Punktzahlen erfasst wurden:

<b>v1</b>	<b>v2</b>	<b>v3</b>
8	10	7
12	8	1
10	8	4
8	10	2
9	9	4

Die Faktorenanalyse wird gleich genauer im Output erklärt.

Nach der Eingabe der Daten müssen im Menü unter dem Button **→Faktorenanalyse** die drei Variablen v1 bis v3 ausgewählt werden. Danach kann man auf Faktorenanalyse klicken. Nun kann man angeben, ab welcher Größe eines Eigenwertes ein Faktor extrahiert werden soll. Standard ist hier 1 („Kaiser-Kreiterium“). Danach kann man auf **→Faktorenanalyse** klicken und man erhält den Output:

## Faktorenanalyse

Die Daten:

	Variable 1	Variable 2	Variable 3
Beobachtung 1	8	10	7
Beobachtung 2	12	8	1
Beobachtung 3	10	8	4
Beobachtung 4	8	10	2
Beobachtung 5	9	9	4

Unten ist die standardisierte Datenmatrix M zu sehen. Wenn Y die Datenmatrix ist, dann erhält man M dadurch, dass man von jeder Spalte j von Y deren Mittelwert

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^n y_{ij}$$

subtrahiert und das Ergebnis durch die (empirische) Standardabweichung der Spalte j

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}$$

geteilt wird:

$$m_{ij} = \frac{y_{ij} - \bar{y}_j}{s_j}$$

Damit ist der Mittelwert der Spalten von M gleich 0 und die (empirische) Varianz gleich 1.

Unten sind die standardisierten Daten (M) zu sehen.

#### Die standardisierten Daten:

	Variable 1	Variable 2	Variable 3
Beobachtung 1	-0.83666002653408	1	1.4768656253944
Beobachtung 2	1.5537971921347	-1	-1.129367831184
Beobachtung 3	0.35856858280032	-1	0.17374889710523
Beobachtung 4	-0.83666002653408	1	-0.69499558842091
Beobachtung 5	-0.23904572186688	0	0.17374889710523

Für die nächsten Berechnungen benötigen wir eine Matrix  $Z$ , für die  $R = Z^T Z$  gilt, wobei  $R$  die (empirische) Korrelationsmatrix ist, die die Korrelationskoeffizienten nach Pearson der Spalten von  $Y$  enthält. Diese Matrix  $Z$  kann man über die Matrix  $M$  bestimmen:  
 $Z = 1/\sqrt{n-1} \cdot M$

**Die (empirische) Korrelationsmatrix R:**

	Variable 1	Variable 2	Variable 3
Variable 1	1	-0.8964214570008	-0.59705025139884
Variable 2	-0.8964214570008	1	0.43437224276307
Variable 3	-0.59705025139884	0.43437224276307	1

Wie zu sehen ist, sind die Nebendiagonalelemente der Korrelationsmatrix  $R$  vom Betrag her recht groß. Man kann also davon ausgehen, dass die Ausgangsvariablen bivariat korreliert sind (was man natürlich mit einem Test zur bivariaten Korrelation untersuchen könnte). Hier wäre also eine Faktorenanalyse durchaus angebracht. Außerdem ist zu sehen, dass die erste Spalte der Datenmatrix mit den beiden anderen negativ korreliert, während die zweite mit der dritten Spalte positiv korreliert ist.

Mit Hilfe der empirischen Korrelationsmatrix  $R$  kann nun die Ladungsmatrix  $L$  und mit ihr die Faktorenmatrix  $F$  berechnet werden. Da  $R = Z^T Z$  gilt, folgt mit der Modellgleichung  $Z = F L^T$  für  $R$ :

$$R = (F L^T)^T F L^T = L F^T F L^T.$$

Wegen der vorausgesetzten Orthogonalität von  $F$  gilt:

$$R = L L^T \quad (1).$$

Da  $R$  zumindest positiv semidefinit und somit diagonalähnlich ist, kann  $R$  wie folgt zerlegt werden:

$$R = T D_i T^T \quad (2),$$

wobei  $T$  die Matrix ist, deren Spalten aus den orthonormierten Eigenvektoren von  $R$  bestehen ( $T$  ist demnach auch eine orthogonale Matrix, für die gilt  $T^T T = I$ ) und  $D$  ist die Diagonalmatrix, die auf der Hauptdiagonalen die (reellen) Eigenwerte von  $R$  enthält.

Somit ergibt sich die gesuchte Ladungsmatrix  $L = T D^{1/2}$ , die der Bedingung (1) genügt.

Mit Hilfe der Eigenwerte der empirischen Korrelationsmatrix  $R$  können wir den Anteil der Varianz bestimmen, die der jeweilige Faktor (in Bezug auf die Gesamtvarianz der Ausgangsvariablen, bzw. der Datenmatrix  $Y$ ) erklärt. Dieser Anteil wurde in % berechnet:

Also erklärt der erste Faktor 76,8563% der Varianz der drei Variablen, der zweite 20,3975% und der dritte 2,74622%.

Sollen nun die Anzahl der Faktoren von  $k$  ( $= p = 3$ ) auf  $k < p$  reduziert werden, so werden nur die Faktoren extrahiert, die möglichst viel Varianz erklären. Hierzu gibt es bestimmte Kriterien, wie z.B. das Kaiserkriterium, nach dem nur die Faktoren extrahiert werden, bei denen der entsprechende Eigenwert größer oder gleich 1 ist, denn nur dadurch wird mehr Varianz erklärt als durch eine der Ausgangsvariablen.

Ein anderes Kriterium wählt so viele Faktoren aus, bis ein willkürlich festgesetzter Anteil an Varianz durch sie erklärt wird. Es kann zur Bestimmung der Faktorenanzahl  $k$  auch der Sphären-Test verwendet werden.

### Eigenwerte von R:

Eigenwert 1	Eigenwert 2	Eigenwert 3
2.3056885692632 (Anteil: 76.86%)	0.08238657702074 (Anteil: 2.75%)	0.61192485371602 (Anteil: 20.4%)

Mit Hilfe der Ladungsmatrix kann die (empirische) Korrelation der einzelnen Faktoren mit den Ausgangsvariablen bestimmt werden. Der erste Faktor korreliert in diesem Modell mit der ersten Variable hoch positiv (0,963485), mit der zweiten Variable hoch negativ (-0,908608) und mit der dritten Variable etwas geringer, aber immer noch negativ (-0,742844). Der zweite Faktor korreliert mit den drei Ausgangsvariablen erheblich geringer, wobei er mit der dritten Variable nur noch sehr gering korreliert (0,0518667).

### Die Ladungsmatrix L:

	Faktor 1	Faktor 2	Faktor 3
Variable 1	0.96348469722779	0.2138200042711	0.16117767829754
Variable 2	-0.90860806560519	0.1843296677653	-0.37477187287566
Variable 3	-0.74284398805423	0.051866717418601	0.66745236012461

Es wird 1 Faktor extrahiert, da 1 Eigenwert größer oder gleich 1 ist (mit einem Anteil von 76.86%).

Nun berechnen wir noch die Matrix „Kommunalität“, auf deren Hauptdiagonalen die Kommunalitäten stehen. Die Kommunalitäten entsprechen der empirischen Korrelation aller Faktoren mit der jeweiligen Ausgangsvariablen. Diese kann auch als Anteil der Varianz definiert werden, die die gemeinsamen Faktoren im Verhältnis zur Gesamtvarianz einer Ausgangsvariablen erklären. Dabei entspricht das erste Diagonalelement dem Anteil der Varianz, die die gemeinsamen Faktoren an der Varianz der ersten Variable erklären u.s.w..

### Die Kommunalitätsmatrix K:

	Variable 1	Variable 2	Variable 3
Variable 1	0.92830276179213	-0.87542996698835	-0.71571881491791
Variable 2	-0.87542996698835	0.82556861688281	0.6749540390324
Variable 3	-0.71571881491791	0.6749540390324	0.55181719058831

Wenn man die Faktorenanalyse mit allen Faktoren durchführen würde, so würde unser Modell keine Fehlermatrix E enthalten und es würden sich auf der Hauptdiagonalen der Kommunalitätenmatrix nur Einsen befinden. Mit dieser Matrix wird nämlich die empirische Korrelationsmatrix R der Daten Y (bzw. die empirische Varianz-Kovarianzmatrix der standardisierten Datenmatrix Y) wie folgt zerlegt:

$$R = \text{Kommunalitätenmatrix} + \text{Residualmatrix}$$

Wenn wie unten nicht alle Faktoren extrahieren werden, so wird man natürlich daran interessiert sein, dass die Hauptdiagonalelemente der Kommunalitätenmatrix möglichst groß sind, so dass der Anteil der erklärten Varianz möglichst groß ist und die Residualmatrix möglichst geringe Hauptdiagonalelemente besitzt.

Betrachtet man die Hauptdiagonale der Matrix Kommunalitäten oben, so stehen hier noch relativ große Werte. Demnach erklärt der eine extrahierte Faktor ca. 92,83% der Varianz der ersten Variablen, ca. 82,56% der zweiten und ca. 55,18% der dritten Variablen.

Wenn man mit allen Faktoren rechnen würde und die Ladungsmatrix damit vollständig wäre, dann kann man die Faktorenmatrix über die untere Gleichung (3) bestimmen, denn aus der Modellgleichung  $Z = F L^T$  folgt:

$$F = Z (L^T)^{-1} \quad (3)$$

Es gilt aber auch  $F = Z T D^{-1/2} = Z T D^{-1} T^T T D^{1/2}$ , also:

$$F = Z R^{-1} L \quad (4)$$

Hiermit kann man die Faktorenmatrix bestimmen, wenn die Ladungsmatrix nicht alle Spalten enthält, wenn nur bestimmte Faktoren extrahiert werden.

In unserem Beispiel ist nur ein Eigenwert größer als eins, womit nur ein Faktor extrahiert wird.

Von den meisten Statistikprogramm Paketen wird zusätzlich die Matrix  $(L^T)^{-1}$  bzw.  $R^{-1}L$  ausgegeben und mit Scores bezeichnet, da über diese mit Hilfe der standardisierten Datenmatrix  $Z$  direkt die Faktorenmatrix  $F$  berechnet werden kann (siehe (4)). Diese enthält also die Linearkombination, über die mit den Spalten der Matrix  $Z$  die Faktorenmatrix  $F$  berechnet werden kann. Da in unserem Beispiel nur der erste Faktor extrahiert wird, wird nur die erste Spalte der Score-Matrix ausgegeben:

#### Die Score-Matrix:

	Faktor 1
Variable 1	0.41787286889992
Variable 2	-0.39407232950612
Variable 3	-0.32217880504634

#### Die Residualmatrix R-K:

	Variable 1	Variable 2	Variable 3
Variable 1	0.071697238207881	-0.020991490012453	0.11866856351907
Variable 2	-0.020991490012453	0.17443138311719	-0.24058179626933
Variable 3	0.11866856351907	-0.24058179626933	0.44818280941169

**Test auf Sphärizität von Bartlett:**

Prüfgröße: 4.6628815724

p-Wert (H0: Korrelationsmatrix ist Einheitsmatrix; gegen H1:

Korrelationsmatrix ist ungleich der Einheitsmatrix): 0.1982 (df = 3)

Die Nullhypothese, dass die Korrelationsmatrix eine Einheitsmatrix ist, kann man auf einem Signifikanzniveau von 5% nicht verwerfen.

Würde man zwei Faktoren extrahieren, dann würden diese beiden Faktoren (mit den größten Eigenwerten) ca. 76,86% + 20,40% = 97,26 % der Varianz der Daten erklären:

Es werden 2 Faktoren extrahiert, da 2 Eigenwerte größer oder gleich 0.6 sind (mit einem Anteil von 97.25%).

Betrachtet man die Hauptdiagonale der Matrix Kommunalitäten, so stehen hier relativ große Werte. Demnach erklären die zwei Faktoren ca. 95,43% der Varianz der ersten Variablen, ca. 96,60% der zweiten und ca. 99,73% der dritten Variablen. Das Modell würde damit die Daten relativ gut erklären.

**Die Kommunalitätsmatrix K:**

	Variable 1	Variable 2	Variable 3
Variable 1	0.95428100577351	-0.93583482734966	-0.60814039313881
Variable 2	-0.93583482734966	0.96602257358154	0.42481166797322
Variable 3	-0.60814039313881	0.42481166797322	0.99730984362422

**Die Score-Matrix:**

	Faktor 1	Faktor 3
Variable 1	0.41787286889992	0.26339456114384
Variable 2	-0.39407232950612	-0.61244754253695
Variable 3	-0.32217880504634	1.0907423616991

### Die Residualmatrix R-K:

	Variable 1	Variable 2	Variable 3
Variable 1	0.045718994226495	0.039413370348867	0.011090141739973
Variable 2	0.039413370348867	0.033977426418465	0.0095605747898467
Variable 3	0.011090141739973	0.0095605747898467	0.0026901563757811

### Test auf Sphärizität von Bartlett:

Prüfgröße: 4.6628815724

p-Wert (H0: Korrelationsmatrix ist Einheitsmatrix; gegen H1: Korrelationsmatrix ist ungleich der Einheitsmatrix): 0.1982 (df = 3)

### Umsetzung mit SAS:

```
Data dat1;
input v1 v2 v3;
datalines;
```

8	10	7
12	8	1
10	8	4
8	10	2
9	9	4

```
run;
```

```
proc factor data = dat1;
var v1 v2 v3;
run;
```

### SAS-Output zur Prozedur FACTOR:

Die Prozedur FACTOR	
Einlesedatentyp	Rohdaten
Anzahl gelesener Datensätze	5
Anzahl verwendeter Datensät	5
N für Signifikanztests	5

Die Prozedur FACTOR  
 Ursprgl. Faktormethode: Hauptkomponenten  
 Priori Kommunalitätsschätzwerte: ONE

<b>Eigenwerte der Korrelationsmatrix: Gesamt = 3</b>				
<b>Durchschn. = 1</b>				
	<b>Eigenwert</b>	<b>Differenz</b>	<b>Proportion</b>	<b>Kumulativ</b>
<b>1</b>	2.30568857	1.69376372	0.7686	0.7686
<b>2</b>	0.61192485	0.52953828	0.2040	0.9725
<b>3</b>	0.08238658		0.0275	1.0000

1 factor will be retained by the MINEIGEN criterion.

<b>Faktormuster</b>	
<b>Factor1</b>	
<b>v1</b>	-0.96348
<b>v2</b>	0.90861
<b>v3</b>	0.74284

<b>Varianz erklärt nach jedem Faktor</b>	
<b>Factor1</b>	
	2.3056886

<b>Endgült. Kommunalität Schätzung: Gesamte = 2.305689</b>			
	<b>v1</b>	<b>v2</b>	<b>v3</b>
	0.92830276	0.82556862	0.55181719



