Multivariate Analysemethoden

Thema: Faktorenanalyse

Günter Meinhardt
Johannes Gutenberg Universität Mainz

Faktorenanalyse

Ziele

- Auffinden latenter Variablen, die einem Set von konkreten Messvariablen zugrunde liegen.
- Reduktion der Messvariablen auf wenige zugrundeliegende orthogonale Dimensionen
- Beurteilung der Konfiguration (Korrelationsstruktur) von Messvariablen durch Anordnung im dem Raum, der durch (wenige) latente Variablen aufgespannt wird.
- Beurteilung der Güte der Lösung:
  - Vollständigkeit der Erklärung der Variablen durch die Faktorlösung
  - Varianzaufklärung durch die einzelnen Faktoren
  - Rotation zu besserer Interpretierbarkeit

Verwendung

Test- und Fragebogenkonstruktion, Screening von Variablenstrukturen, Vorbereiten von Klassifikationsanalysen, Testen von Hypothesen und Voraussetzungen, Testen von Pfadmodellen, Konfirmatorische Analysen von Hypothesensets

Latente Variable

Faktorenanalyse

Man möchte Messvariablen in einem Raum latenter Dimensionen (Fähigkeiten, Traits) anordnen. Gegeben ist ein Set von Beobachtungen (Messvariablen)

$$\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\right)$$

Problem: Finde latente Variablen

$$\left(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r\right)$$

$r \leq m$, so dass jede Variable $\mathbf{x}_k$ eine Linearkombination der $\mathbf{w}_l$ ist:

$$\mathbf{x}_k = b_{k1}\mathbf{w}_1 + b_{k2}\mathbf{w}_2 + \ldots + b_{kr}\mathbf{w}_r$$

Beispiel:
Das Abschneiden im Abitur mit Deutsch, Mathe, Physik, Latein und Geographie wird erklärt aus latenten Variablen Memory, Induction, Perceptual Speed, Space, Verbal Comprehension.

Beispiel

Life Satisfaction

Work

X_1: Gehalt
X_2: Entscheidungsfreiheit
X_3: Qualität der Kommunikation

Privacy

X_4: Ehe
X_5: Freunde/Beziehungen
X_6: Sexualität

Person

X_7: Lebensansprüche
X_8: Sinnhaftigkeit

Activity

X_9: Hobbies
X_{10}: Sport/Fitness

$$\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{10}\right)$$

$m = 10$ Messvariablen, erhoben an $n = 100$ Testpersonen

Ausgangspunkt

Korrelationsmatrix R

	Geh	Ents	Com	Ehe	Freu	Sex	Anspr	Sinn	Hob	Fit
Gehalt	**1.00**	0.65	0.65	0.60	0.52	0.14	0.15	0.14	0.61	0.55
Entscheid	0.65	**1.00**	0.73	0.69	0.70	0.14	0.18	0.24	0.71	0.68
QualComm	0.65	0.73	**1.00**	0.64	0.63	0.16	0.24	0.25	0.70	0.67
Ehe	0.60	0.69	0.64	**1.00**	0.80	0.54	0.63	0.58	0.90	0.84
Freunde	0.52	0.70	0.63	0.80	**1.00**	0.51	0.50	0.48	0.81	0.76
Sex	0.14	0.14	0.16	0.54	0.51	**1.00**	0.66	0.59	0.50	0.42
Anspruch	0.15	0.18	0.24	0.63	0.50	0.66	**1.00**	0.73	0.64	0.59
Sinn	0.14	0.24	0.25	0.58	0.48	0.59	0.73	**1.00**	0.59	0.52
Hobby	0.61	0.71	0.70	0.90	0.81	0.50	0.64	0.59	**1.00**	0.84
Fitness	0.55	0.68	0.67	0.84	0.76	0.42	0.59	0.52	0.84	**1.00**

$m \times m$ symmetrische Korrelationsmatrix wird faktorisiert

⬇

Principal Components Analysis (PCA)

1. Schritt

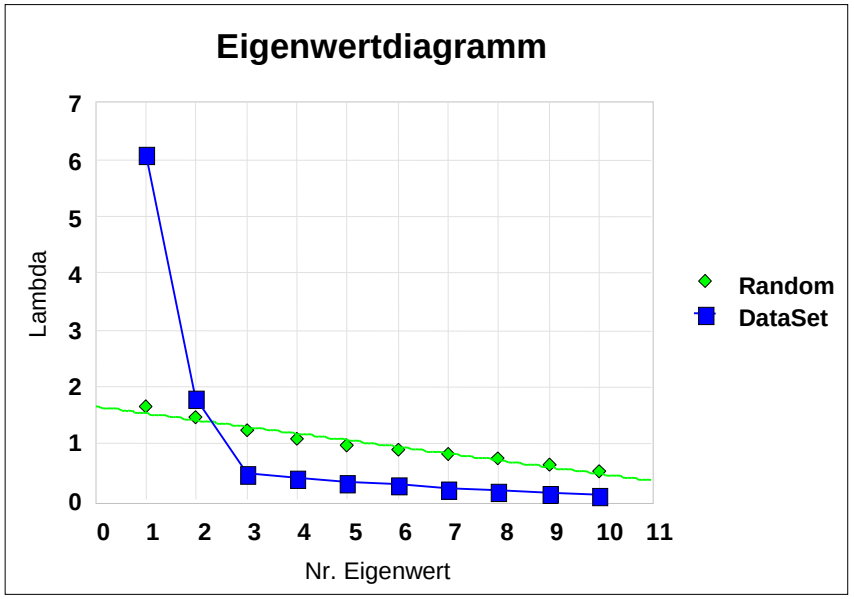
Ladungsmatrix B

	Fac 1	Fac 2	Fac 3	Fac 4	Fac 5	Fac 6	Fac 7	Fac 8	Fac 9	Fac 10	h^2
Gehalt	-0.653	0.514	0.302	0.439	-0.014	0.127	0.051	-0.022	0.080	0.004	1
Entscheid	-0.757	0.495	-0.079	-0.212	-0.091	0.172	-0.236	0.163	0.104	0.012	1
QualComm	-0.746	0.457	-0.105	0.031	-0.205	-0.422	0.034	0.019	-0.018	0.039	1
Ehe	-0.942	-0.022	0.013	0.002	0.121	0.094	-0.024	0.002	-0.243	0.172	1
Freunde	-0.876	0.052	0.100	-0.325	-0.016	0.091	0.313	-0.026	0.089	0.018	1
Sex	-0.576	-0.605	0.491	-0.115	-0.113	-0.115	-0.145	-0.024	0.004	-0.020	1
Anspruch	-0.671	-0.618	-0.126	0.160	0.225	-0.104	0.028	0.201	0.145	0.048	1
Sinn	-0.642	-0.574	-0.269	0.153	-0.363	0.160	0.011	-0.080	0.007	0.001	1
Hobby	-0.952	0.014	-0.050	0.027	0.077	0.013	0.036	0.096	-0.157	-0.224	1
Fitness	-0.900	0.048	-0.152	-0.035	0.227	-0.051	-0.121	-0.293	0.088	-0.030	1
Expl.Var	6.118	1.801	0.473	0.408	0.317	0.293	0.196	0.170	0.138	0.085	10
Prp.Totl	0.612	0.180	0.047	0.041	0.032	0.029	0.020	0.017	0.014	0.009	1

zeigt die Koordinaten der m Variablen auf den m Faktoren

2. Schritt

Anzahl der Faktoren ermitteln



Factor Scree Test (Cattell):

- Normalverteilte Zufallszahlen ziehen für dieselbe Anzahl von Variablen mit derselben Anzahl von Messungen (VPn)
- Die Anzahl der Eigenwerte über der Scree-Test Geraden ist die Anzahl von Faktoren, die als interpretierbar angesehen wird (Horn-Kriterium).

➡ hier: 2 Faktor Lösung

3. Schritt (a)

Ladungsmatrix

Güte der 2 Faktor Lösung

	Factor 1	Factor 2	h^2
Gehalt	-0.653	0.514	0.690
Entscheid	-0.757	0.495	0.818
QualComm	-0.746	0.457	0.765
Ehe	-0.942	-0.022	0.887
Freunde	-0.876	0.052	0.769
Sex	-0.576	-0.605	0.698
Anspruch	-0.671	-0.618	0.833
Sinn	-0.642	-0.574	0.741
Hobby	-0.952	0.014	0.906
Fitness	-0.900	0.048	0.813
Expl. Var	6.118	1.801	
Prop.Total	0.612	0.180	

Güte der Lösung

- Zeile: zeigt die Güte der Varianzaufklärung jeder Variable (Kommualität)
- Spalte: zeigt die Güte der Varianzaufklärung jedes Faktors (NormEigen)

Summe der Ladungsquadrate läßt die Güte der Faktorlösung beurteilen

3. Schritt (b)

Vergleich
$R - R_h$

Korrelationsmatrix R

	Geh	Ents	Com	Ehe	Freu	Sex	Anspr	Sinn	Hob	Fit
Gehalt	**1.00**	0.65	0.65	0.60	0.52	0.14	0.15	0.14	0.61	0.55
Entscheid	0.65	**1.00**	0.73	0.69	0.70	0.14	0.18	0.24	0.71	0.68
QualComm	0.65	0.73	**1.00**	0.64	0.63	0.16	0.24	0.25	0.70	0.67
Ehe	0.60	0.69	0.64	**1.00**	0.80	0.54	0.63	0.58	0.90	0.84
Freunde	0.52	0.70	0.63	0.80	**1.00**	0.51	0.50	0.48	0.81	0.76
Sex	0.14	0.14	0.16	0.54	0.51	**1.00**	0.66	0.59	0.50	0.42
Anspruch	0.15	0.18	0.24	0.63	0.50	0.66	**1.00**	0.73	0.64	0.59
Sinn	0.14	0.24	0.25	0.58	0.48	0.59	0.73	**1.00**	0.59	0.52
Hobby	0.61	0.71	0.70	0.90	0.81	0.50	0.64	0.59	**1.00**	0.84
Fitness	0.55	0.68	0.67	0.84	0.76	0.42	0.59	0.52	0.84	**1.00**

3. Schritt (b)

Vergleich
$R - R_h$

Reproduzierte Korrelationsmatrix R_h

	Geh	Ents	Com	Ehe	Freu	Sex	Anspr	Sinn	Hob	Fit
Gehalt	**0.69**	0.75	0.72	0.60	0.60	0.06	0.12	0.12	0.63	0.61
Entscheid	0.75	**0.82**	0.79	0.70	0.69	0.14	0.20	0.20	0.73	0.71
QualComm	0.72	0.79	**0.76**	0.69	0.68	0.15	0.22	0.22	0.72	0.69
Ehe	0.60	0.70	0.69	**0.89**	0.82	0.56	0.65	0.62	0.90	0.85
Freunde	0.60	0.69	0.68	0.82	**0.77**	0.47	0.56	0.53	0.83	0.79
Sex	0.06	0.14	0.15	0.56	0.47	**0.70**	0.76	0.72	0.54	0.49
Anspruch	0.12	0.20	0.22	0.65	0.56	0.76	**0.83**	0.79	0.63	0.57
Sinn	0.12	0.20	0.22	0.62	0.53	0.72	0.79	**0.74**	0.60	0.55
Hobby	0.63	0.73	0.72	0.90	0.83	0.54	0.63	0.60	**0.91**	0.86
Fitness	0.61	0.71	0.69	0.85	0.79	0.49	0.57	0.55	0.86	**0.81**

Residualkorrelationen: $R_e = R - R_h$

3. Schritt (b)

Vergleich
$R_e = R - R_h$

Residualkorrelationen R_e

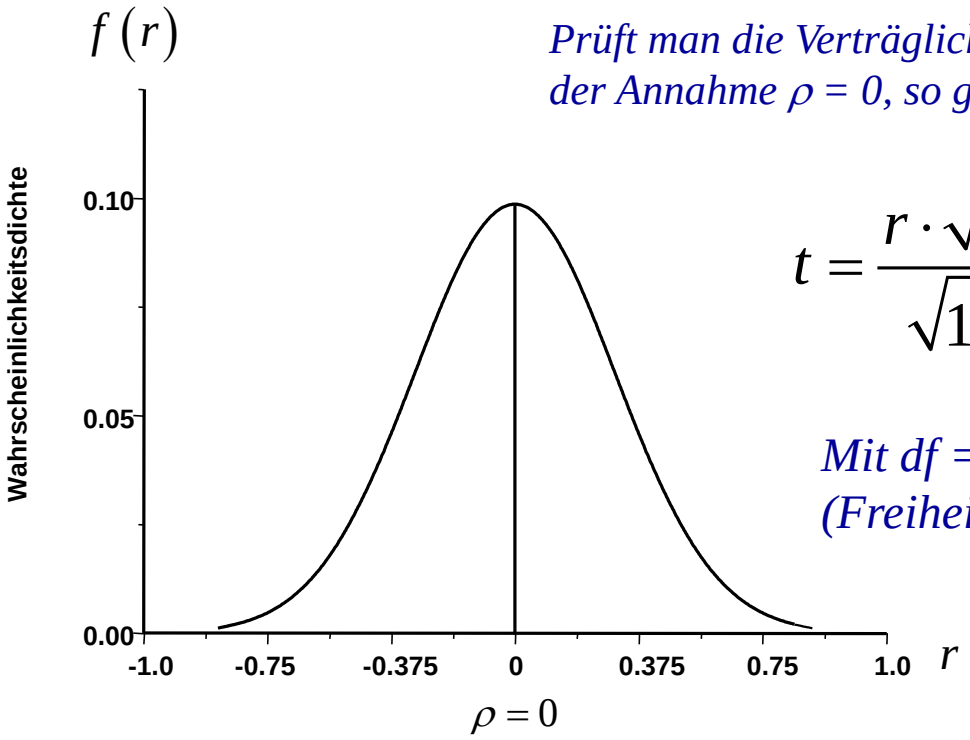
	Geh	Ents	Com	Ehe	Freu	Sex	Anspr	Sinn	Hob	Fit
Gehalt	0.31	-0.10	-0.07	-0.01	-0.08	0.08	0.02	0.01	-0.02	-0.06
Entscheid	**-0.10**	0.18	-0.06	-0.01	0.01	0.01	-0.02	0.03	-0.02	-0.02
QualComm	**-0.07**	**-0.06**	0.24	-0.06	-0.05	0.01	0.02	0.04	-0.02	-0.02
Ehe	**-0.01**	**-0.01**	**-0.06**	0.11	-0.02	-0.02	-0.01	-0.03	0.01	-0.00
Freunde	**-0.08**	**0.01**	**-0.05**	**-0.02**	0.23	0.03	-0.06	-0.05	-0.02	-0.04
Sex	**0.08**	**0.01**	**0.01**	**-0.02**	**0.03**	0.30	-0.10	-0.13	-0.04	-0.06
Anspruch	**0.02**	**-0.02**	**0.02**	**-0.01**	**-0.06**	**-0.10**	0.17	-0.05	0.01	0.02
Sinn	**0.01**	**0.03**	**0.04**	**-0.03**	**-0.05**	**-0.13**	**-0.05**	0.26	-0.02	-0.03
Hobby	**-0.02**	**-0.02**	**-0.02**	**0.01**	**-0.02**	**-0.04**	**0.01**	**-0.02**	0.09	-0.02
Fitness	**-0.06**	**-0.02**	**-0.02**	**-0.00**	**-0.04**	**-0.06**	**0.02**	**-0.03**	**-0.02**	0.19

Analyse der Residualkorrelationen:
Prüfung der Verträglichkeit mit Null - Korrelation

3. Schritt (b)

Test der r_e

Testen der Verteilung der Residualkorrelationen



Prüft man die Verträglichkeit von r mit der Annahme $\rho = 0$, so gilt:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Mit df = n – 2 (Freiheitsgrade)

Die Verteilung von Korrelationskoeffizienten um den Erwartungswert $\rho = 0$ folgt einer t- Verteilung.

➡ Man bestimmt die kritischen Grenzen für r_e

3. Schritt (b)

Test der r_e

Testen der Verteilung der Residualkorrelationen

$$r_0 = \pm \frac{t_0}{\sqrt{n-2+t_0^2}} = \pm \frac{1.98}{\sqrt{98+1.98^2}} = \pm 0.197$$

Mit $t_0 = t_{(.975;98)}$

Vergleich mit kritischen Korrelationen zeigt:

Keine $r_e > r_0$

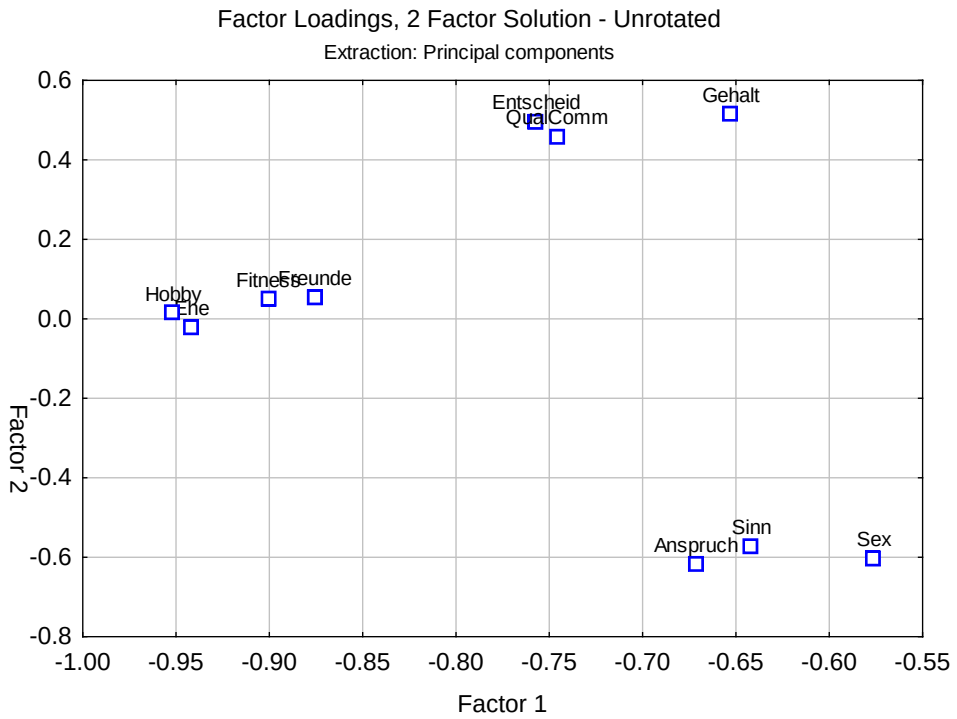
Residualkorrelationen sind reine Zufallskorrelationen, (spurious correlations), die keine systematischen Beziehungen über Restfaktoren enthalten.

Die exakte Übereinstimmung mit der t- Verteilung kann Mit einem Verteilungs-Anpassungstest geprüft werden.

4. Schritt

Darstellung

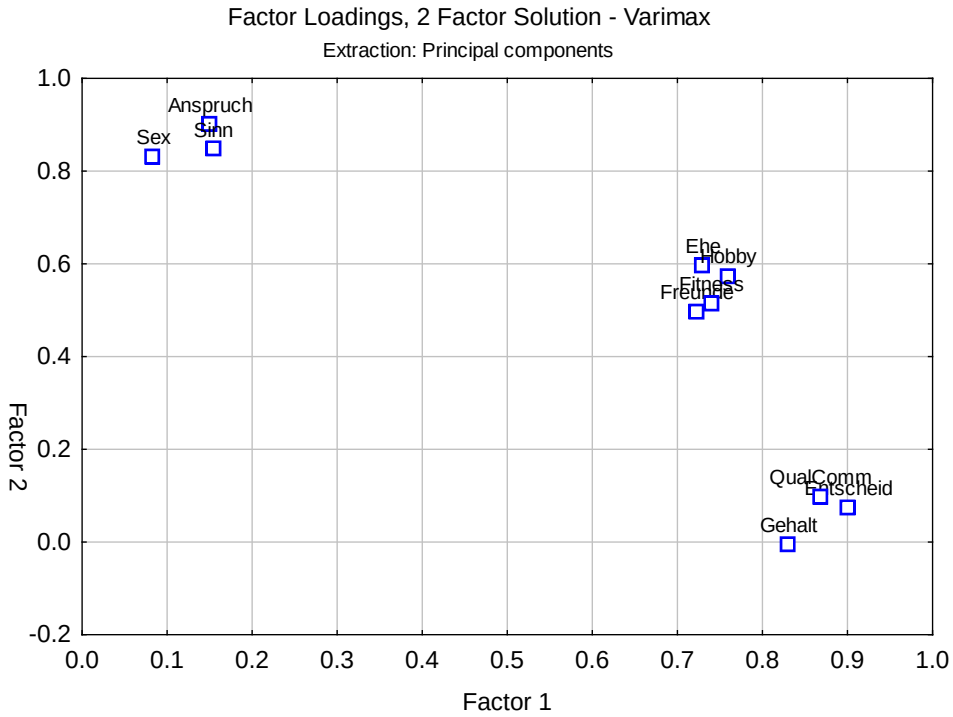
Variablen im Faktorenraum



- 3 Variablencluster
- Lösung wird rotiert zu besserer Interpretierbarkeit

4. Schritt

Varimax – rotierte Lösung



Interpretation

- 2 unabhängige Lebensbereiche & eine Mischform
- Lösung wird rotiert zu besserer Interpretierbarkeit

4. Schritt

Ladungsmatrix

Vergleich: rotiert - unrotiert

rotiert

	Factor 1	Factor 2	h^2
Gehalt	0.831	-0.006	0.690
Entscheid	0.901	0.074	0.818
QualComm	0.869	0.097	0.765
Ehe	0.730	0.595	0.887
Freunde	0.723	0.496	0.769
Sex	0.084	0.831	0.698
Anspruch	0.151	0.900	0.833
Sinn	0.155	0.847	0.741
Hobby	0.760	0.573	0.906
Fitness	0.741	0.514	0.813
Expl.Var	4.493	3.426	7.919
Prp.Totl	0.449	0.343	0.792

unrotiert

	Factor 1	Factor 2	h^2
Gehalt	-0.653	0.514	0.690
Entscheid	-0.757	0.495	0.818
QualComm	-0.746	0.457	0.765
Ehe	-0.942	-0.022	0.887
Freunde	-0.876	0.052	0.769
Sex	-0.576	-0.605	0.698
Anspruch	-0.671	-0.618	0.833
Sinn	-0.642	-0.574	0.741
Hobby	-0.952	0.014	0.906
Fitness	-0.900	0.048	0.813
Expl.Var	6.118	1.801	7.919
Prp.Totl	0.612	0.180	0.792

- Kommunalitäten bleiben konstant
- Gesamtvarianzaufklärung bleibt konstant
- Varianzanteile der Faktoren ändern sich

5. Schritt

Faktorwerte

Faktorwerte für rotierte Lösung

Da jede Variable $\mathbf{z}_k$ eine Linearkombination der $\mathbf{F}_l$ ist:

$$\mathbf{z}_k = b_{k1}\mathbf{F}_1 + b_{k2}\mathbf{F}_2 + \ldots + b_{kr}\mathbf{F}_r$$

muss die (n × r) Matrix der Werte der latenten Variablen geschätzt werden.

Mit weniger Faktoren als Variablen (hier: 2) gilt nur approximativ

$$\mathbf{z}_k \approx b_{k1}\mathbf{F}_1 + b_{k2}\mathbf{F}_2$$

Für das Problem gibt es 2 Methoden

- multiple Regressionsschätzungen
- Schätzungen über die Diagonalmatrix der Eigenwerte

für hohe Kommunalitäten sind beide Methoden äquivalent, für geringe sind multiple Regressionsschätzungen vorzuziehen.