Aufgabenserie 5: Hauptkomponentenanalyse, Faktorenanalyse

Aufgabe 11

Folgende Tabelle enthält die Daten von zwölf amerikanischen Städten, für welche zehn Variablen erhoben wurden. Während sich die Variablen X_2 bis X_7 auf die Luftverschmutzung beziehen, sind die übrigen Variablen demografischer Natur, wobei X_1 die "Sterblichkeitsrate", X_8 die "Bevölkerungsdichte pro Quadratmeile mal 0.1", X_9 den "Anteil an Weißen in der Bevölkerung" und X_{10} den "Anteil an Familien, die ein Einkommen oberhalb der Armutsgrenze beziehen" bezeichnet.[1] Die Variablen sollen letztlich zur Prognose der zukünftigen Zu- und Abwanderung dienen; zunächst ist aber die Anzahl der exogenen Variablen mithilfe einer Hauptkomponentenanalyse zu reduzieren. Die Daten liegen Ihnen in der Datei `staedte.txt` vor.

Stadt	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
A	1199	155	229	340	63	147	253	1357.2	93.1	87.3
B	841	2	61	188	54	126	229	25.4	95.8	86.9
C	921	65	134	236	49	150	299	150.2	94	90.4
D	869	18	27	128	22	122	754	28.6	69	73.7
E	1112	42	163	337	55	141	252	174.5	97.3	88.5
F	938	137	205	308	32	91	182	103.3	94.7	90.7
G	1000	75	166	328	88	182	296	167.5	85.2	89.4
H	689	40	46	58	10	78	157	20.9	87,2	75,2
I	938	1	47	179	32	69	141	26.2	95.2	88.8
J	823	47	67	248	29	129	284	25.3	67.7	74.6
K	823	31	46	158	28	66	142	15.2	70.2	67.8
L	780	15	283	940	55	225	958	27.9	94.2	78.6

Führen Sie für die gegebenen Daten mittels `R` eine Hauptkomponentenanalyse durch:

1. Berechnen Sie die Stichproben-Varianz-Kovarianz- und die Stichproben-Korrelationsmatrix.
2. Geben Sie an, ob Sie die Analyse auf Basis der Stichproben-Varianz-Kovarianz-Matrix oder auf Basis der Stichproben-Korrelationsmatrix durchführen würden. Begründen Sie Ihre Entscheidung.
3. Bestimmen Sie anhand der Eigenwerte der gewählten Matrix die Anzahl der zu bildenden Hauptkomponenten.
4. Berechnen Sie die erste Hauptkomponente.
5. Zeigen Sie allgemein, dass die (normierten) Eigenvektoren von $\mathbf{\Sigma}$ das Optimierungsproblem der Hauptkomponentenanalyse, d. h. $\max_{\boldsymbol{l}} \boldsymbol{l}'\mathbf{\Sigma}\boldsymbol{l}$ unter den Nebenbedingungen $\boldsymbol{l}'\boldsymbol{l} = 1$ und Unkorreliertheit der Hauptkomponenten, lösen.

[1] Die Daten sind Jobson, J.D.: Applied Multivariate Data Analysis, Springer, 1992, S. 702 entnommen.

Aufgabe 12

Zu einer Korrelationsmatrix $\boldsymbol{\rho}$ von vier Zufallsvariablen X_1 bis X_4 liegt Ihnen der Vektor $\boldsymbol{\lambda}$ ihrer Eigenwerte und die aus den korrespondierenden Eigenvektoren gebildete Matrix $\boldsymbol{P}$ vor:

$$\boldsymbol{\lambda} = \begin{pmatrix} 1.875 \\ 1.421 \\ ? \\ 0.239 \end{pmatrix}, \quad \boldsymbol{P} = \begin{pmatrix} -0.605 & 0.365 & -0.139 & 0.694 \\ -0.621 & 0.331 & 0.034 & -0.709 \\ -0.308 & -0.648 & -0.693 & -0.066 \\ -0.391 & -0.580 & 0.707 & 0.105 \end{pmatrix}.$$

1. Inwiefern unterscheidet sich die Faktorenanalyse in ihrem Ansatz und ihrer Zielsetzung von der Hauptkomponentenanalyse?

2. Bestimmen Sie den fehlenden Eigenwert λ_3 und die Korrelationsmatrix $\boldsymbol{\rho}$. Berechnen Sie sodann die Determinante und die Spur von $\boldsymbol{\rho}$.

3. Extrahieren Sie die ersten beiden Faktoren mittels der Hauptkomponenten-Methode, und stellen Sie die Ladungsmatrix bei Verwendung dieser beiden Faktoren auf.

4. Bestimmen Sie die Kommunalitäten basierend auf den beiden Faktoren. Interpretieren Sie das Element l_{21} der Ladungsmatrix.

5. Welcher Prozentsatz der Streuung der Variablen X_3 kann durch die beiden gebildeten Faktoren erklärt werden?

6. Welcher Prozentsatz der gesamten Streuungssumme wird vom ersten Faktor erklärt?

Aufgabe 13

Ein Online-Versandhändler interessiert sich dafür, was Kunden zur Abgabe von Produktbewertungen motiviert. Um dies untersuchen, wurden 100 Bewerter befragt. Auf einer Siebener-Skala von "stimme voll zu" bis "stimme überhaupt nicht zu" sollten sie jeweils angeben, inwieweit sie ein bestimmtes Motiv mit der Abgabe einer Bewertung verfolgen:

X_1: "Ich möchte andere Kunden vor einem schlechten Produkt warnen."
X_2: "Ich möchte anderen Kunden ein gutes Produkt empfehlen."
X_3: "Ich möchte eine Belohnung (z. B. einen Warengutschein) erhalten."
X_4: "Ich möchte einen Mehrwert für die Kunden-Community schaffen."
X_5: "Ich möchte mir eine gute Reputation als Produktbewerter aufbauen."

Folgende Tabelle enthält die Stichproben-Korrelationsmatrix, die sich aus der Befragung ergab:

	X_1	X_2	X_3	X_4	X_5
X_1	1.000	0.902	-0.054	0.713	0.049
X_2	0.902	1.000	-0.086	0.798	0.063
X_3	-0.054	-0.086	1.000	0.012	0.862
X_4	0.713	0.798	0.012	1.000	0.111
X_5	0.049	0.063	0.862	0.111	1.000

Diese Stichproben-Korrelationsmatrix liegt Ihnen auch in der Datei `korrelationen.txt` vor. Auf ihrer Basis soll eine Faktorenanalyse durchgeführt werden.

1. Diskutieren Sie, ob die verwendeten Daten aus messtheoretischer Sicht überhaupt für eine Faktorenanalyse geeignet sind.

2. Wann ist im Rahmen der Faktorenanalyse die Annahme der Normalverteilung nötig?

3. Testen Sie mithilfe des Bartlett-Tests die Hypothese, dass den beobachteten Variablen ein gemeinsamer Faktor zugrunde liegt.

4. Überprüfen Sie, ob gemäß des Kriteriums kleiner Restkorrelationen die Verwendung eines Faktors akzeptabel wäre.

Gehen Sie nun von zwei zu extrahierenden Faktoren aus.

5. Bestimmen Sie die Ladungsmatrix zum einen mittels der Hauptkomponenten-Methode und zum anderen mithilfe der Maximum-Likelihood-Methode. Stellen Sie die Ergebnisse vergleichend gegenüber.

6. Berechnen Sie für beide Methoden die Kommunalitäten und den erklärten Streuungsanteil je Faktor. Vergleichen Sie die Ergebnisse.

7. Führen Sie für die mittels der Maximum-Likelihood-Methode extrahierten Faktoren eine Faktor-Rotation durch, und interpretieren Sie die rotierten Faktoren.