h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 7

A 1 In einer Studie wurde bei 100 Personen jeweils die Expression von 1000 Genen gemessen. Die Daten sind in einer 100x1000-Matrix zusammengefasst. Die Probanden sind in zwei Gruppen eingeteilt, erkrankte Personen (K) und gesunde Personen (G). Die Fragestellung ist, welche der Gene in den beiden Gruppen unterschiedlich exprimiert werden. Dazu wird als Prä-Analyse zunächst eine PCA durchgeführt mit dem Ergebnis, dass die erste Hauptkomponente "10% der Variation erklärt".

Erläutern Sie, was mit "10% der Variation erklärt" gemeint ist.

A 2 Es sei X eine 6-dimensionale Zufallsvariable mit Mittelwertsvektor und Kovarianzmatrix

$$\mu = (3.0, 2.5, 7.1, 5.1, 0.7, 9.2),$$

$$\Sigma = \begin{pmatrix} 1 & 0.2 & 0.3 & 0.4 & 0.4 & 0.05 \\ 0.2 & 1 & 0.45 & 0.3 & 0.15 & 0.2 \\ 0.3 & 0.45 & 1 & 0.1 & 0.2 & 0.4 \\ 0.4 & 0.3 & 0.1 & 1 & 0.15 & 0.4 \\ 0.4 & 0.15 & 0.2 & 0.15 & 1 & 0.05 \\ 0.05 & 0.2 & 0.4 & 0.4 & 0.05 & 1 \end{pmatrix}.$$

Die Eigenwerte der Kovarianzmatrix sind (gerundet):

$$\Lambda = (0.96,\ 1.16,\ 2.27,\ 0.75,\ 0.27,\ 0.60).$$

Die zugehörigen normierten Eigenvektoren der Kovarianzmatrix sind (gerundet):

$$\begin{pmatrix} -0.13 \\ 0.34 \\ 0.55 \\ -0.69 \\ 0.16 \\ -0.25 \end{pmatrix}, \begin{pmatrix} 0.51 \\ -0.20 \\ -0.23 \\ -0.04 \\ 0.59 \\ -0.55 \end{pmatrix}, \begin{pmatrix} -0.43 \\ -0.43 \\ -0.46 \\ -0.43 \\ -0.32 \\ -0.38 \end{pmatrix}, \begin{pmatrix} -0.09 \\ -0.66 \\ 0.20 \\ -0.25 \\ 0.42 \\ 0.53 \end{pmatrix}, \begin{pmatrix} 0.42 \\ 0.30 \\ -0.51 \\ -0.51 \\ -0.08 \\ 0.45 \end{pmatrix}, \begin{pmatrix} 0.60 \\ -0.38 \\ 0.38 \\ -0.10 \\ -0.59 \\ -0.09 \end{pmatrix}.$$

a) Bestimmen Sie für eine Beobachtung $X = (2, 1, 8, 6, 0, 7)$ die ersten beiden Hauptkomponenten Y_1 und Y_2.
b) Wie viel Varianz erklären die ersten beiden Hauptkomponenten?
c) Zeichnen Sie einen screeplot.

A 3

a) Schauen Sie sich den Datensatz `USArrests` im R-Paket `datasets` inkl. der Variablenbeschreibungen an
b) Schauen Sie sich die paarweisen Scatterplots für die `USArrests`-Daten an.
c) Führen Sie eine Hauptkomponentenanalyse für die Daten im Datensatz `USArrest` durch. Sie können dazu z.B. die R-Funktionen `princomp`, `screeplot` und `biplot` nutzen. Interpretieren Sie das Ergebnis:
   - Wie viel Varianz wird durch die verschiedenen Hauptkomponenten erklärt? Erstellen Sie einen Screeplot.
   - Erstellen Sie einen Plot der beiden ersten Hauptkomponenten und interpretieren Sie diesen.
   - Welche Variablen sind positiv/negativ mit den beiden Hauptkomponenten korreliert? In welcher Graphik und woran ist dieser Zusammenhang zu erkennen?
d) Führen Sie die gleichen Analysen nach Standardisierung der Variablen durch. Interpretieren Sie die Unterschiede in den Ergebnissen. Welches Vorgehen wäre Ihrer Meinung nach sinnvoller und warum?