Hauptkomponentenanalyse

Aufgabe 1:
Laden Sie den Datensatz `wetter.txt` sowie die zugehörige Variablenbeschreibung von der Übungshomepage herunter. Importieren Sie den Datensatz in R.

a) Bestimmen Sie Erwartungswertvektor, Kovarianz- und Korrelationsmatrix der Daten.

b) Machen Sie sich mit der Funktion `princomp()` vertraut. Bestimmen Sie den Anteil der Gesamtvarianz der einzelnen Hauptkomponenten bzgl. der Korrelationsmatrix. Stellen Sie die Varianzen graphisch mit Hilfe von `screeplot()` dar. Berechnen Sie nun mit Hilfe der Funktion `loadings()` die Eigenvektoren bzgl. der Korrelationsmatrix.

c) Wie b) bzgl. der Kovarianzmatrix.

d) Interpretieren Sie die Hauptkomponenten bzgl. der Kovarianzmatrix und die Hauptkomponenten bzgl. der Korrelationsmatrix. Welche Interpretation ist sinnvoller?

Aufgabe 2: Der Datensatz `europa.txt` enthält für 24 europäische Länder die Variablen Oberfläche (`ober`), Einwohnerzahl (`einw`), Bruttosozialprodukt (`brut`) und Arbeitslosigkeit (`arbl`) (vgl. Blatt 7)

a) Die Kovarianzmatrix sieht folgendermaßen aus:

```
              ober          einw          brut         arbl
ober 32978148684.3 2702249.50290 -10555022.464 225450.14058
einw     2702249.5     538.70080     -7754.047     23.36583
brut   -10555022.5   -7754.04710 119364491.123 -34229.79529
arbl      225450.1      23.36583    -34229.795     20.67346
```

Warum ist es nicht sinnvoll, die Hauptkomponentenanalyse ohne Skalierung der vier Variablen, d.h. auf der Basis der Kovarianzmatrix durchzuführen?

b) Die Korrelationsmatrix, das heisst die Kovarianzmatrix $\mathbf{\Sigma}_{scaled}$, die sich ergibt, wenn man vorher die vier Variablen auf Varianz 1 skaliert, sieht folgendermaßen aus:

```
             ober        einw         brut       arbl
ober  1.000000000  0.64111878 -0.005319961  0.2730428
einw  0.641118783  1.00000000 -0.030578561  0.2214119
brut -0.005319961 -0.03057856  1.000000000 -0.6890649
arbl  0.273042838  0.22141186 -0.689064920  1.0000000
```

Die R-Funktion `eigen()` berechnet die Eigenwerte und Eigenvektoren einer Matrix. Für $\mathbf{\Sigma}_{scaled}$ liefert sie folgendes Ergebnis:

```
$values [1] 1.9465661 1.4235518 0.3731246 0.2567576

$vectors
           [,1]       [,2]       [,3]       [,4]
[1,] -0.4959586 -0.4927052  0.6173002 -0.3608423
[2,] -0.4842364 -0.4975888 -0.7066954  0.1360225
[3,]  0.4361680 -0.5963492  0.2420757  0.6289074
[4,] -0.5738445  0.3924470  0.2468229  0.6751046
```

Wie groß ist der Anteil der Gesamtstreuung, der durch die ersten beiden Hauptkomponenten zusammen erklärt wird?

c) Wieviele Hauptkomponenten halten Sie für notwendig?

d) Interpretieren Sie die vier Hauptkomponenten mit Hilfe von vier Balkendiagrammen, mit denen Sie das Gewicht jeder ursprünglichen Variable in jeder Hauptkomponente darstellen.

e) Wie werden für jedes Land die Werte, die ihm durch die ersten beiden Hauptkomponenten zugewiesen werden, berechnet? Berechnen Sie diese für die Schweiz und Bulgarien. Interpretieren Sie die Ergebnisse.