

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2010
Blatt 11

Aufgabe 11.1 (5 Punkte)

Die folgende Aufgabe zum Thema Hauptkomponentenanalyse können Sie sowohl mit Bleistift und Papier als auch mithilfe von R lösen.

- a) Berechnen Sie die Loadings g_1 und g_2 für die Kovarianzmatrix

$$S = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

und geben Sie die Hauptkomponenten z_1 und z_2 an.

Berechnen Sie außerdem den Anteil der Gesamtvarianz, der durch die erste Hauptkomponente z_1 erklärt wird.

- b) Transformieren Sie die Kovarianzmatrix S aus Teil a) in eine Korrelationsmatrix R . Berechnen Sie auch hier die Loadings g_1^* und g_2^* sowie die Hauptkomponenten z_1^* und z_2^* . Welcher Anteil der Gesamtvarianz wird durch z_1^* erklärt?
- c) Vergleichen Sie die in den Teilen a) und b) berechneten Hauptkomponenten. Sind sie gleich? Sollten sie gleich sein?

Aufgabe 11.2 (5 Punkte)

Auf der Homepage liegt der bekannte Schweizer-Banknoten-Datensatz `bank.txt`. Er enthält die Ergebnisse von Längenmessungen an 200 Schweizer 1000-Franc-Scheinen (100 echten und 100 gefälschten) in der Einheit mm. Welche Längen genau gemessen wurden, können Sie der Infodatei (`info.txt`) sowie dem Bild in der Datei `franc.pdf` entnehmen.

- a) Führen Sie Hauptkomponentenanalysen für den gesamten Datensatz durch und zwar sowohl
- auf der Basis von Kovarianzen als auch
 - auf der Basis von Korrelationen.

Dies ist in R mithilfe der Funktionen `princomp` und `prcomp` möglich.

- b) Wieviele Hauptkomponenten würden Sie wählen, um eine Dimensionsreduktion durchzuführen? Schauen Sie sich dazu den `screeplot` und den Prozentsatz der erklärten Gesamtvarianz (z. B. mit der Funktion `summary`) an.

Für welche der beiden Hauptkomponentenanalysen aus Teil a) sollten die Loadings überhaupt interpretiert werden? Interpretieren Sie die Loadings der ersten Hauptkomponente.

- c) Erstellen Sie für beide Hauptkomponentenanalysen aus Teil a) einen Biplot (in R mit der Funktion `biplot` möglich). Welche Scores-Struktur liegt vor? Vergleichen und interpretieren Sie die Biplots.