Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

**Fachbereich Informatik**
**Sommersemester 2024**
Prof. Dr. Peter Becker

# Mathematisch-algorithmische Grundlagen für Data Science

## Aufgabenblatt 12

Abgabe zu **zweit** vor der Vorlesung am 2. Juli 2024.

Sollpunktzahl: 0 Punkte

### Aufgabe 1 (Hauptkomponentenanalyse) 10 Punkte

Auf der Homepage der Vorlesung finden Sie den Datensatz `australian-crabs.csv`, der 200 Beobachtungen zu australischen Krabben (Leptograpsus variegatus) enthält. Die acht Attribute des Datensatzes sind:

1. Art (Blue bzw. Orange)
2. Geschlecht (male bzw. female)
3. laufende Nummer innerhalb von Art + Geschlecht
4. Größe des Frontallappens (FL, mm)
5. hintere Breite (RW, mm)
6. Panzerlänge (CL, mm)
7. Panzerbreite (CW, mm)
8. Körpergröße (BD, mm)

Führen Sie eine Hauptkomponentenanalyse für die morphologischen Merkmale (4. bis 8.) dieses Datensatzes durch. Gehen Sie dabei wie folgt vor:

(a) Berechnen die den Mittelwert für jedes Merkmal und zentrieren Sie damit den Datensatz.

(b) Berechnen Sie die Kovarianzmatrix.

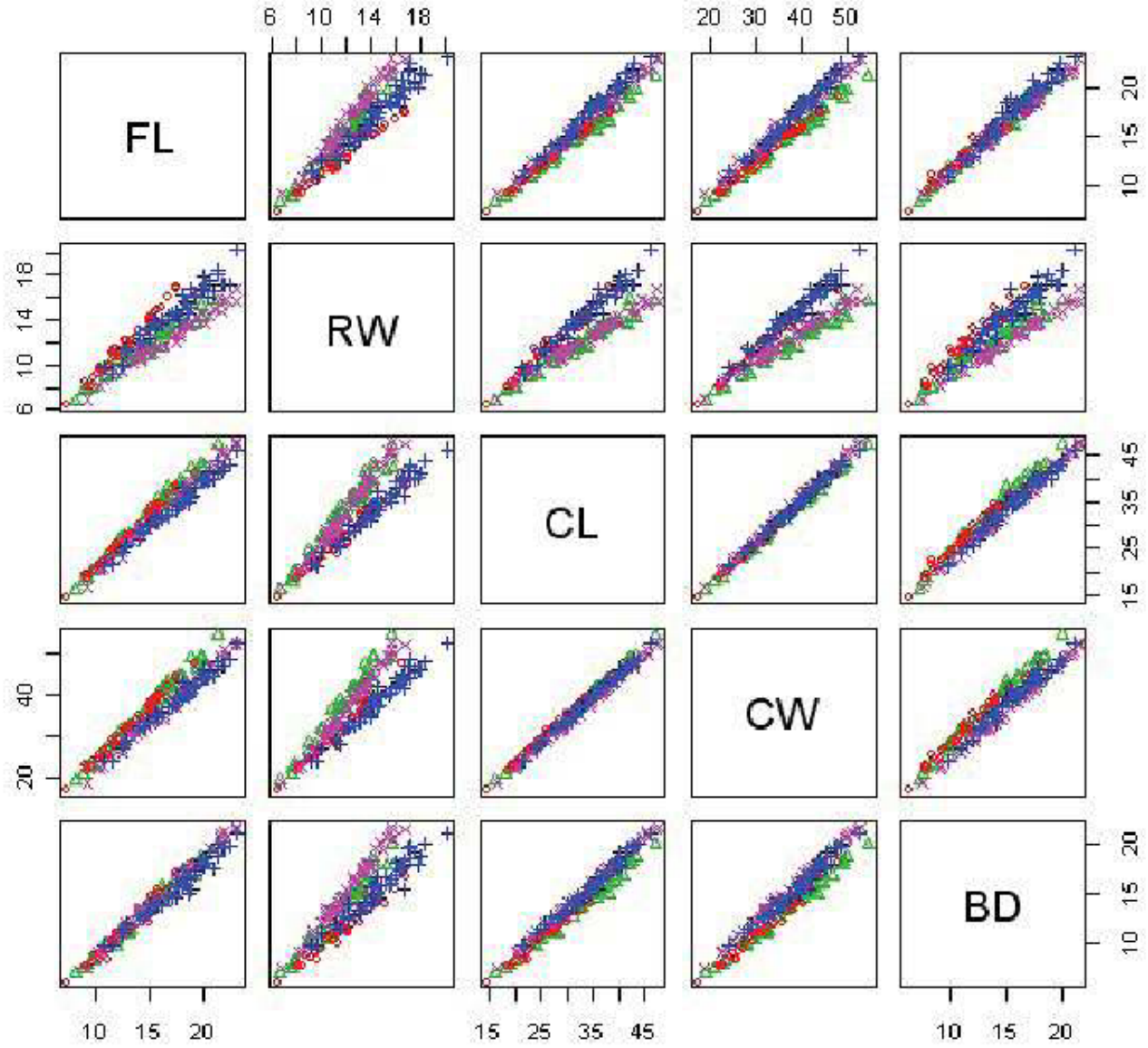
(c) Berechnen Sie die Eigenwerte und Eigenvektoren der Kovarianzmatrix.

(d) Geben Sie die Hauptkomponenten sowie deren Anteil an der Gesamtvarianz an.

(e) Selektieren Sie die erste und die zweite Hauptkomponente und transformieren Sie damit den Datensatz auf zwei Merkmale.

(f) Plotten Sie die transformierten Daten. Stellen Sie dabei die vier Klassen (Blue, male), (Blue, female), (Orange, male) und (Orange, female) durch unterschiedliche Farben dar.

Hier die Scatterplots für jeweils zwei Merkmale:




**Hinweise:**

- Sie können bspw. die Java-Klassen nutzen, die auf der Homepage veröffentlicht sind (siehe unter den Folien von Kapitel 6).
- Für die Berechnung der Eigenwerte und Eigenvektoren nutzen Sie die Klasse `QR`, die wiederum `GS` nutzt.
- Das QR-Verfahren (Methode `QR.qr()`) konvergiert bei mir mit einem `eps`-Wert von $10^{-10}$. Ein Wert von $10^{-8}$ ist aber auch vollkommen ausreichend.