

# Übungsblatt – Principal Component Analysis (PCA)

## Aufgabe 1: Konzeptuelles Verständnis

- a) Benenne Sie ein Problem, das bei hochdimensionalen Daten auftreten kann. Erklären Sie, wie die PCA dabei hilft.
- b) Beschreiben Sie mit eigenen Worten die Funktionsweise der PCA
- c) Warum müssen die Daten vor der PCA zentriert werden? Erläutern Sie, welches Problem auftritt, wenn man diesen Schritt überspringt.
- d) Wie viele PCs wählt man für 90% Varianzerhalt, wenn die Eigenwerte 4, 2.5, 1.8, 0.9, 0.5 und 0.3 sind?

## Aufgabe 2: Kovarianz

Gegeben sei folgender Datensatz:

$a$	4	-6	10	4
$b$	6	-4	8	6

- a) Berechnen Sie die jeweiligen Mittelwerte  $\bar{a}$  und  $\bar{b}$  der Daten. Zentrieren Sie dann die Daten, indem Sie die jeweiligen Mittelwerte von den einzelnen Beobachtungen abziehen.
- b) Die Kovarianz zweier Stichproben  $a$  und  $b$  mit jeweils  $i$  Beobachtungen berechnet sich wie folgt:

$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b}).$$

Berechnen Sie die Kovarianz zwischen  $a$  und  $b$ .

- c) Die Varianz von  $a$  ist:  $\text{var}(a) = 33$ , die Varianz von  $b$  ist:  $\text{var}(b) = 22$ . Geben Sie die Kovarianzmatrix an.
- d) Angenommen die Variable  $a$  hätte invertierte Vorzeichen, also  $a = -4, 6, -10, -4$ . Überlegen Sie was das für die Beziehung zwischen den Variablen bedeutet und welchen Einfluss das auf die Kovarianz hätte.



Alle Materialien zur Veranstaltung finden Sie auf GitHub:  
<https://github.com/MartinaEchtenbruck/Principal-Component-Analysis>