6. The following Python code loads the crabs datasets into a data frame $\mathbf{X}$.

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
from scipy.linalg import eigh # eigendecomposition
from numpy.linalg import svd # SVD

# Load the crabs dataset
url = 'https://vincentarelbundock.github.io/Rdatasets/csv/MASS/crabs.csv'
crabs = pd.read_csv(url)
X = crabs[['FL', 'RW', 'CL', 'CW', 'BD']]
```

Using Python (recommended) or R:

(a) Compute the sample covariance S

(b) Compute the eigenvalues and the principal components using the eigendecomposition of S

(c) Compute the PCA projections, and plot them using a pairplot (function pairplot in seaborn)

(d) Compute the eigenvalues and the principal components using the SVD decomposition of $\mathbf{X}$

(e) Compute the Gram matrix B

(f) Compute the PCA projections from the eigendecomposition of B

François Caron: caron@stats.ox.ac.uk

7. (Optional) Spike sorting is an important problem in computational neuroscience. Based on recordings of neuronal activity, it aims at detecting spikes, and assigning each spike to the activity of a given neuron. The dataset we will consider is composed of $n = 1000$ spike recordings. For each spike, a set of $p = 96$ features have been extracted using some preprocessing tools. The data[1] are stored in the csv file spike_data.csv. Here is the code to load the data as a panda frame.

```
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
X = pd.read_csv('spike_data.csv')
```

Using Python (recommended) or R:

(a) Apply PCA to the data X, and compute the projections of the data onto the first two components.

(b) Apply Kmeans with two clusters to the PCA projections, and visualise the obtained partition.

(c) Let $\widehat{\mu}_1 \in \mathbb{R}^2$ and $\widehat{\mu}_2 \in \mathbb{R}^2$ be the estimated cluster means. Map these means into the original data space of dimension 96 using the PCA decoder. Plot the two transformed cluster means.

[1] This is a subset of the data available from: https://ifcs.boku.ac.at/repository/data/spike_sorting/index.html

François Caron: caron@stats.ox.ac.uk