Principal component analysis with Python and scikit learn

Exercise 1 : PCA of synthetic data

One can use the parts `More on random variables and random vectors` and `Visualization with Python` of the notebook `ComplementsPracticalSession2.ipybn`

1. Generate first 500 3D vectors following a standard Gaussian distribution. Then visualize your synthetic data.

2. Apply PCA to these data. Comment.

3. Generate another dataset and apply PCA to these new data. Explain what happens

4. We define now the following dilation and rotation matrices
   >>> `s1 = np.array([[3,0,0],[0,1,0],[0,0,0.2]])`
   >>> `r1 = np.array([[0.36,0.48,-0.8],[-0.8,0.6,0],[0.48,0.64,0.6]])`
   Apply this dilation and rotation to these data. What happens?

5. Apply a PCA to these observations. Comment

Exercise 2 : PCA on data on mamals

We want to analyze the data `mamals.csv`. To download the data, one can use the part `Importing csv files with Numpy` of the notebook `ComplementsPracticalSession2.ipynb`. One should also import the name of the species using a command of the form
>>> `noms = np.genfromtxt('/path/mammals.csv', dtype='str', delimiter=';', usecols=[0], skip_header=1)`

1. Apply PCA to the data `mammals` and display eigenvalues. What happens?

2. Normalize the data and apply PCA. Sort eigenvalues and plot them. Comment

3. Display projection on the two first axes and thereafter on the three first ones

Exercise 3 : PCA on the data `leaf`

We now apply PCA on data concerning tree leafs coming from
`https://archive.ics.uci.edu/ml/datasets/Leaf`.
To download the data, one can use the part `Importing csv files with Numpy` of the notebook `ComplementsPracticalSession2.ipybn`.

1. Apply PCA to the data and display the eigenvalues :

2. Display the projections of the data on the 3 first principal components. Comment

3. Display the projections of the data on the 3 first principal components giving different colors to the labels. The label is located on the first column of `leaf`.