Exercice 1 *Woven PCA.* The data table X below consists of three variables x, y and z, and four individuals A to D.

$$X = \begin{array}{c|ccc} & x & y & z \\ \hline \text{A} & 1 & 0 & 0 \\ \text{B} & 1 & 2 & 0 \\ \text{C} & 2 & 2 & 2 \\ \text{D} & 0 & 0 & 2 \end{array}$$

In the later, you can compute exact values using hand made computations (but we strongly recommend the library `sympy`). You can also use the package `numpy.linalg`.

1. Compute the center of gravity g of the cloud
2. Center the array X
3. (a) Compute the variance-covariance matrix S of the cloud. Give the variances of the variables. How to interpret a covariance equal to zero ?
   (b) What is the inertia of the cloud ?
4. Find the principal axes of inertia :
   (a) Determine the eigenvalues of S
   (b) Verify your result with the help of question 3b
   (c) Determine the first two eigenvalues (projected variance) and eigenvectors (loadings)
   (d) Compute the ratio of explained inertia by these two axes
5. Find the projection of the individuals on these axes

Bonus : Answer the previous question using the reduced centered array.

Exercice 2 *Breast Cancer data set.* The aim of this work is to compare the result of **logistic** regressions on the *Breast Cancer Wisconsin (Diagnostic)* DataSet using raw data as inputs and principal components as inputs.

The data set can be obtained using

```
from sklearn.datasets import
    load_breast_cancer
cancer = load_breast_cancer()
```

and more informations about it can be obtained at the kaggle site. You can also find on this site a lot of Notebooks related with this data set (but you should try to build your notebook from scratch).

The tasks you have to perform are

1. Represent the original data set in order to understand it. First, separate `maligant` and `benign` data using

```
malignant = cancer.data[cancer.
    target == 0]
```

```
benign = cancer.data[cancer.target
    == 1]
```

and then build the histograms of the 30 input variables

```
from matplotlib.colors import
    ListedColormap
cm3 = ListedColormap(['#0000aa', '#
    ff2020', '#50ff50'])

fig, axes = plt.subplots(15, 2,
    figsize=(10, 20))

ax = axes.ravel()
for i in range(30):
    _, bins = np.histogram(cancer.
        data[:, i], bins=50)
    ax[i].hist(malignant[:, i],
        bins=bins, color=cm3(0),
        alpha=.5)
    ax[i].hist(benign[:, i], bins=
        bins, color=cm3(2), alpha=.5)
    ax[i].set_title(cancer.
        feature_names[i])
    ax[i].set_yticks(())
ax[0].set_xlabel("Feature magnitude
    ")
ax[0].set_ylabel("Frequency")
ax[0].legend(["malignant", "benign"
    ], loc="best")
fig.tight_layout()
```

Interpret the histograms.

2. Perform a logistic regression in order to predict the cancer output variable. Evaluate the results either using a test sample or using directly the learning sample.
3. Find meaningful representations of the transformed and reduced data set in a low dimensional space using PCA. interpret the first two components.
4. Perform a logistic regression using the components as input. How many components should you retain ?

Exercice 3 *Wild images Data Set.* The aim of this work is to show how PCA can be used in order to perform features extraction. The last particular part allow you to learn the Non-Negative Matrix Factorization method (NME).

Read and follow the instructions in the notebook (may be time consuming).