6 Principal Component Analysis

Exercise 6.1
Let us consider the data matrix X with 5 samples of dimension 2 :

$$X = \begin{pmatrix} 0 & 17 \\ 2 & 19 \\ 3 & 21 \\ 4 & 23 \\ 6 & 20 \end{pmatrix}.$$

1. Calculate the matrix Y obtained by centering and scaling the matrix X.
2. Calculate the covariance matrix Σ_Y of Y.
3. Calculate by hand the PCA of Y.
4. Describe the best rank one approximation of Y. Draw the cloud of points given by Y and the best PCA line of Y. Deduce the equation of the best PCA line which approximates the data in X.
5. Use Python and the SVD to compute the PCA of X. Plot with Matplotlib the cloud of points X and the best PCA line of X.

Exercise 6.2
1. Load the data set "turtles.csv" with Pandas. Transform the dataframe into an array.
2. Use the function "Axes3D.scatter" to plot the numerical data in 3D. The colors of the samples must depend on the sex of the turtles (red triangle for male and blue circle for female). Do not forget the labels of the axes. What do you conclude from the visual aspect of the data set ?

   Hints : consult the tutorial http ://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html
3. Use the command "sklearn.preprocessing.scale" to center and scale the data. Use again "Axes3D.scatter" to show the result.
4. Use the library "sklearn.decomposition.PCA" to calculate the full PCA.
5. From the result of the PCA, plot the explained variance per principal component. Plot also the screeplot. You must change the "xticks" with the labels "PC1", "PC2" and "PC3".
6. Which variables did contribute the most in Principal Component 1 (PC1) ? Is it possible to give a meaning to PC1 ?
7. Choose a relevant number of components and compute the PCA. Plot in 3D the scaled dataset and the approximated samples (samples projected on the PCA subspace) on the same figure. The projected samples must be plotted with yellow squares. Is the PCA approximation relevant ?