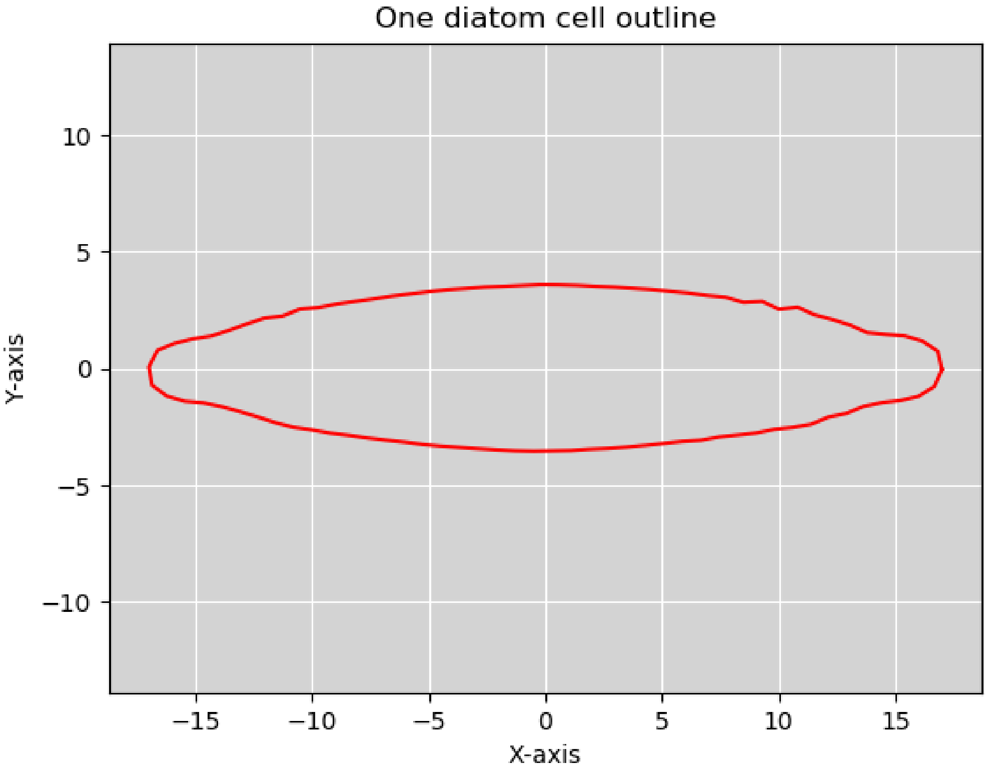
Stefano Pellegrini - Assignment 4

March 12, 2020

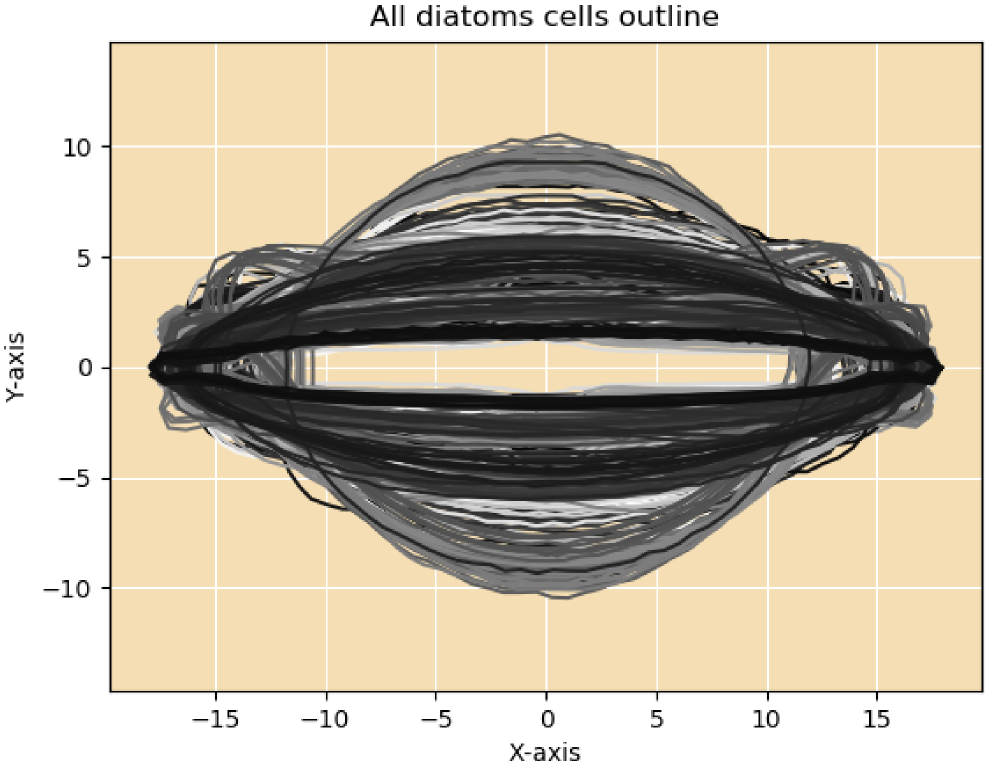
Data exploration with PCA

Exercise 1

- **Plot of a cell shape**



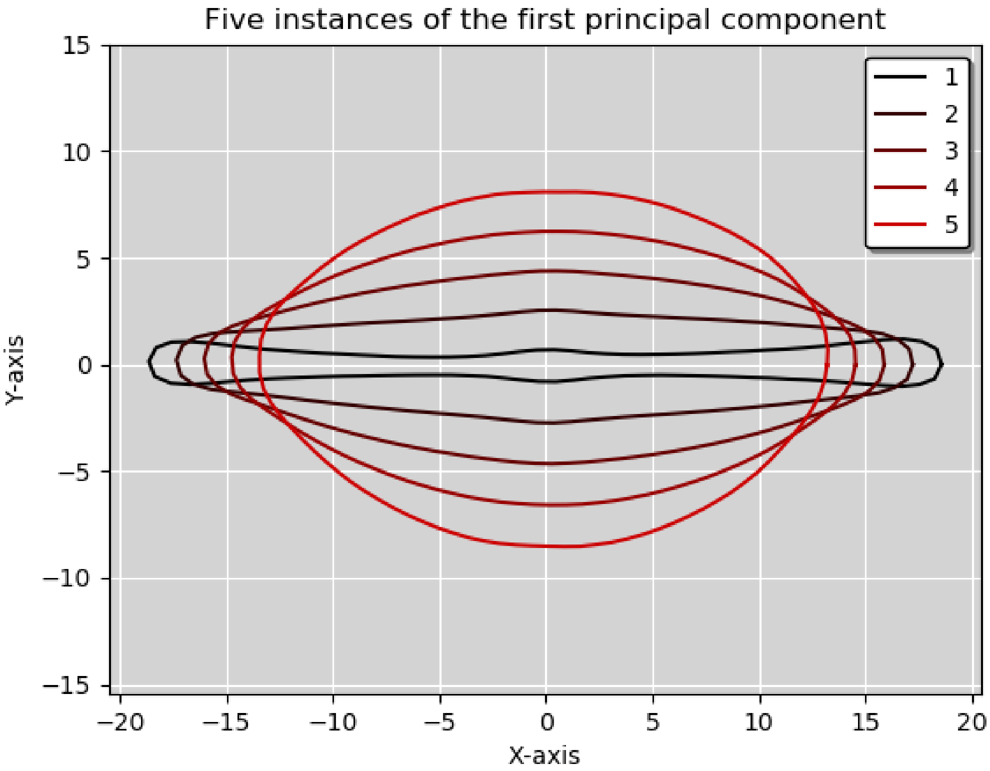
- **Plot of many cells**



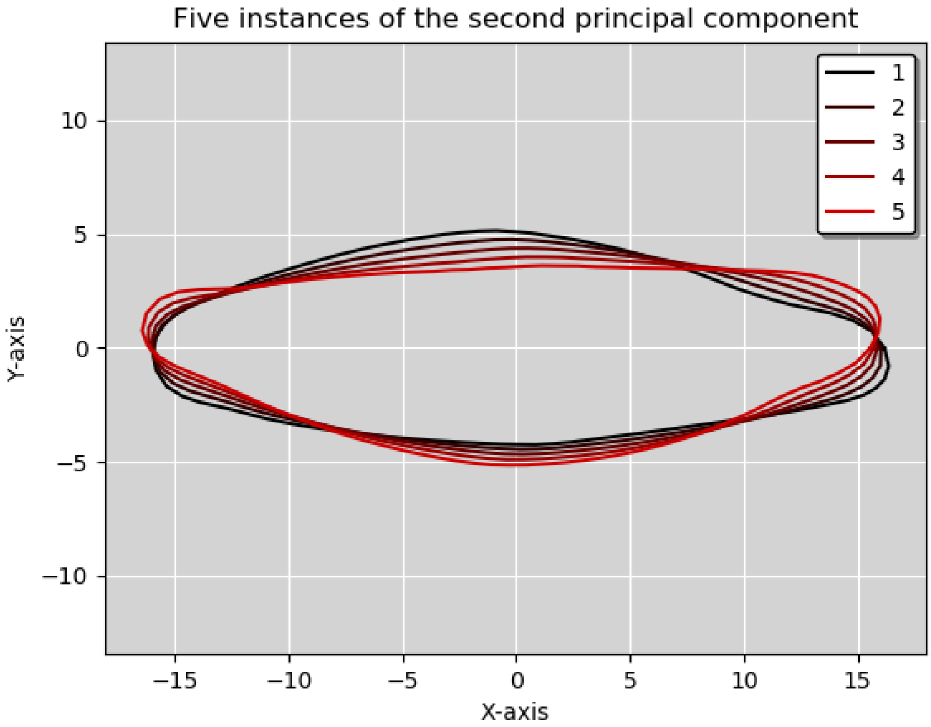
- **Description**
  The dataset describe the outline of different diatom cells. In the plot of all diatoms is possible to observe that the 780 cells have similar shapes. Most of them are elliptical, some being more stretched than others, while some other show bilateral simmetry. Overal the outline of the cells show a quite similar pattern.

Exercise 2

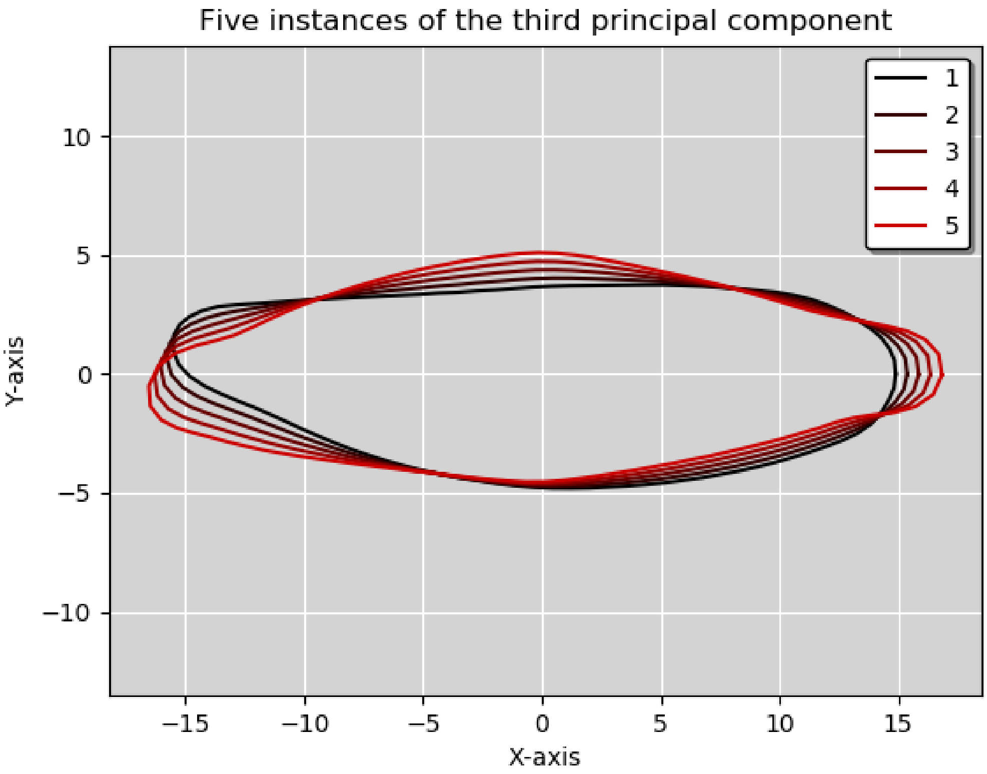
- **Plot of first PC instances**



- **Plot of second PC instances**



- **Plot of third PC instances**



- **Description of the three components**
  The first component captures the largest amount of variance (77%) in the shapes of the diatom outline. It describes the different stretch of the cells, ranging from very shrank or contracted shapes to almost sferical one.
  The second and third components capture respectively 15% and 2% of the variance, these two components describe minor differences in the outline shape of the cells.

Exercise 3

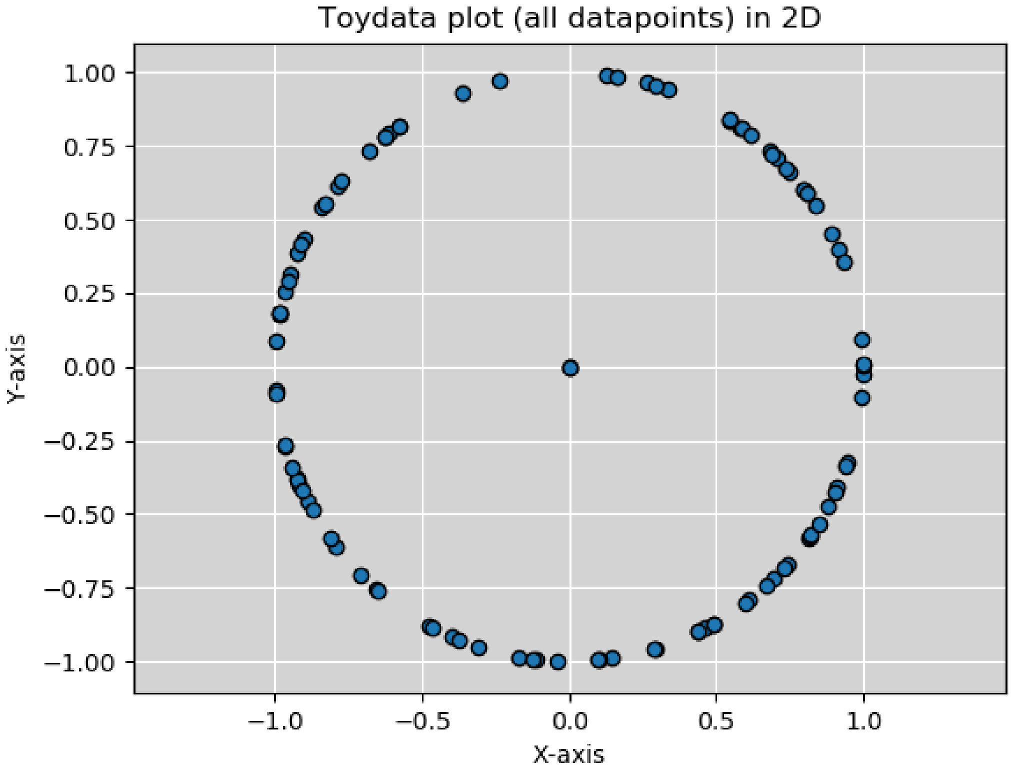
- Exercise 3a

PCA is a procedure that project the data into new dimensions, defined by the principal components (eigenvectors), in such a way that the variance of the projected data points is maximized.

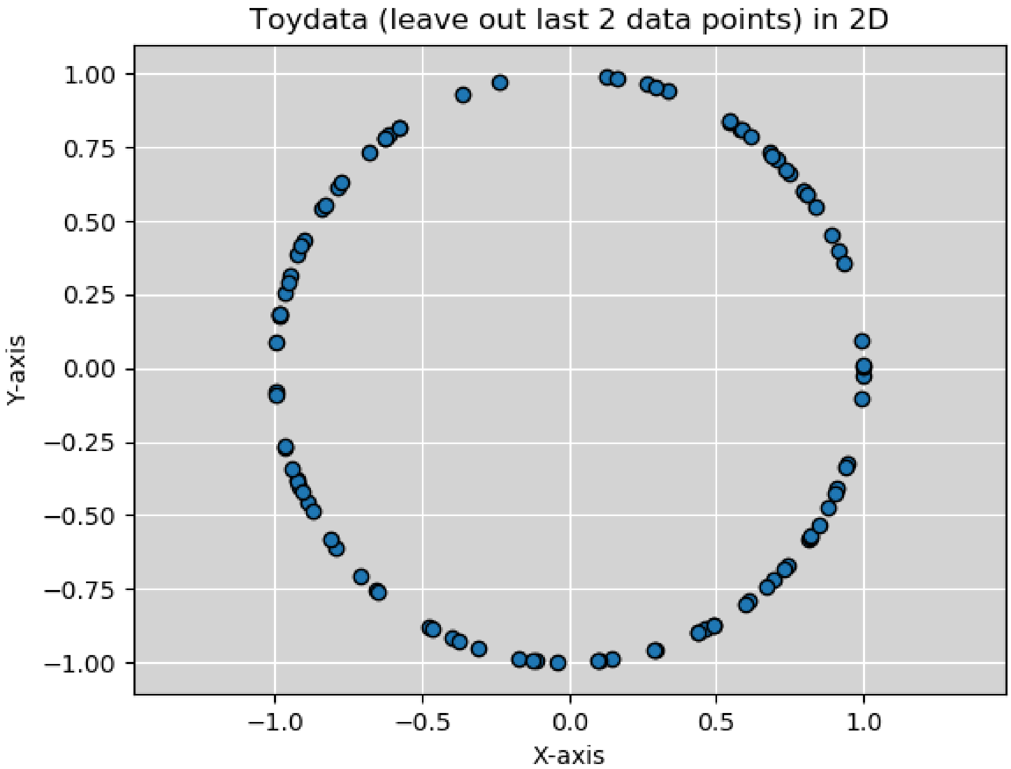
1. Centering will move the data points to the center of the axis, unaffecting the variance because it doesn't change the distances between the data points. Centering the data prior to PCA is often a good idea because it may be helpful for downstream analysis.
2. Standardizing means that the data is centered and also normalized. Rescaling the data prior to PCA may be usefull if the features are not in the same scale, but it can be a bad idea otherwise because we will lose part of the original variance.
3. Whitening means that the data is centered, decorrelated and then normalized. I don't think that performing whitening prior to PCA is usefull because we will lose most of the variance of the data.

- Exercise 3b

- **Plot of toydata projection**



- **Plot of toydata projection leaving out the last two data points**



- **Explaination**
  The last two points correspond to the center of the axis. They are located close to the mean of the data points, but if we consider the underline distribution of the data, they could be considered outliers because they are not in continuity with the rest of the points. So removing the two last points can be seen as removing the outliers from the dataset.

Clustering II

Exercise 4

- **Description of software used**
  I started by centering the dataset. In order to obtain the centroids I used my own implementation of the unsupervised k-means clustering algorithm. I initialized the algorithm placing the 2 centroids at the 2 first points of the dataset. To evaluate the quality of the clusters I used my implemented loss function that computes the sum of the squared errors among them. Based on the distances between the points and the centroids, I assign each point to the nearest cluster. Then I update the new centroids computing the mean of the new assigned points, and finally, I compute the loss of the new clusters. The algorithm continues this process until the centroids converge, which occurs when the loss also converge. Once I obtained the centroids I used them to obtain the data points assigned at each clusters. Then I performed MDS (using my own implemented function) to project the data in reduced dimensions. After that I divide the dataset according to their classes and to the obtained clusters, then I obtained the centroids in the reduced dimensional space by

computing the mean of the projected clusters (in 2D or 3D). Finally I plotted the projected data points (using different colors according to their classes) and the projected cluster centers.

- **Projection of the two cluster centers in 2D, centered pesticide training data**

  - $\underline{The\, first\, centroid\,(2D)}$ = (-1596.80176436, -123.75257989)

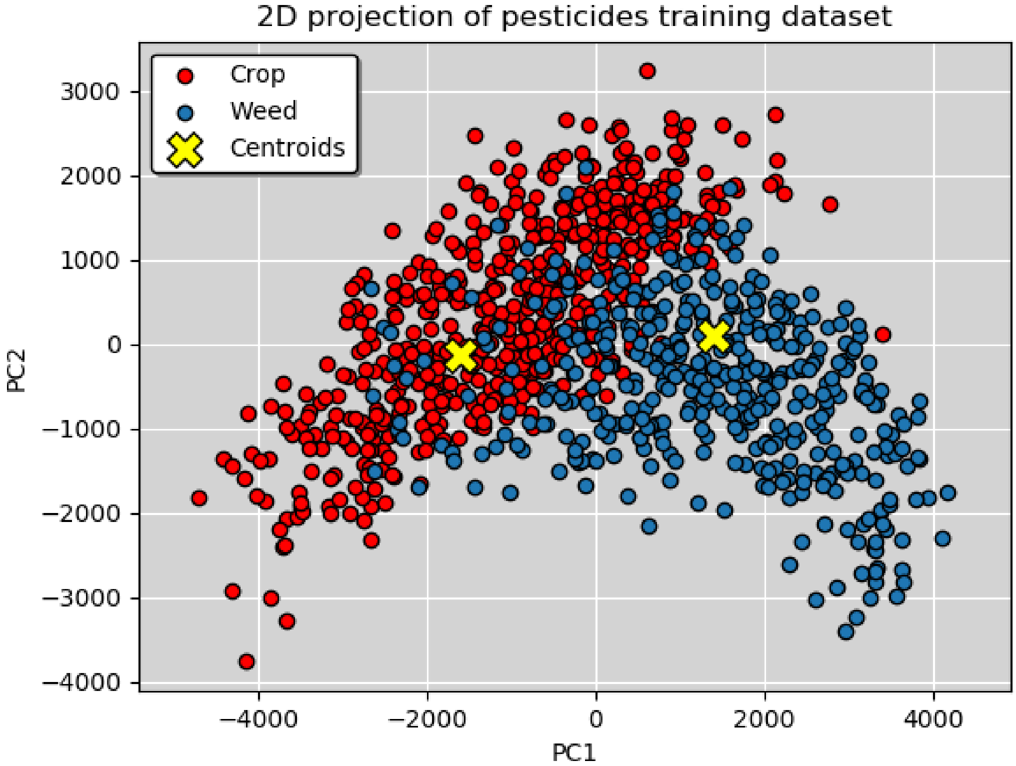
  - $\underline{The\, second\, centroid\,(2D)}$ = (1404.7053115, 108.86505148)

- **Projection of the two cluster centers in 3D, centered pesticide training data (not requested)**

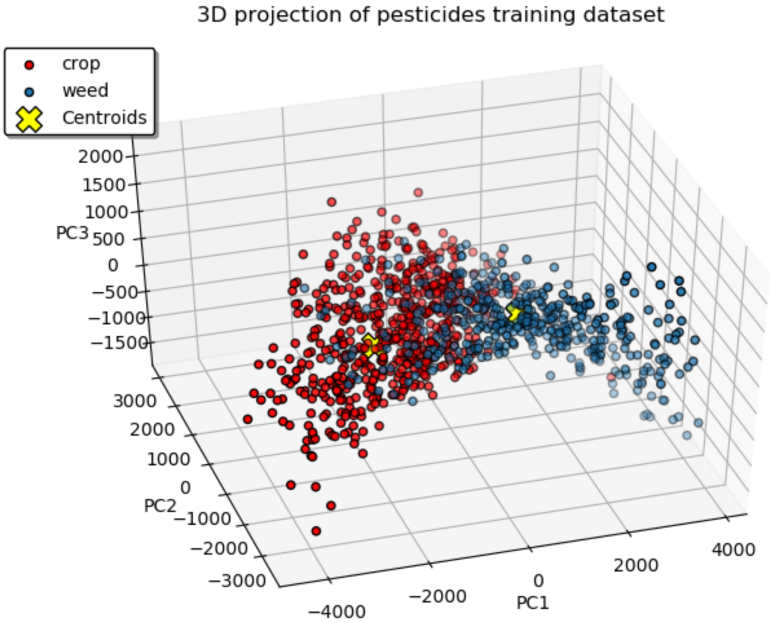
  - $\underline{The\, first\, centroid\,(3D)}$ = (-1596.80176436, -123.75257989, -27.2294927)

  - $\underline{The\, second\, centroid\,(3D)}$ = (1404.7053115, 108.86505148, 23.95376425)

- **2D plot visualizing the data and the cluster centers**



- **3D plot visualizing the data and the cluster centers (not requested)**



- **Short discussion of the results**
  The cluster centers obtained with the k-means algorithm seem to be quite reasonable. Despite that, the classes division visualized in 2D and 3D doesn't seem to show a clear clusters separations, indeed it seems to be more continuous.