# IDS2018 - Assignment 4

Isabela Blucher

March 22, 2018

## Exercise 1 (Plotting cell shapes)

Figure 1 represents the first cell on the diatoms dataset. The landmark points were plotted with the scatter function, and the lines between the points with the normal plot function from matplotlib.pyplot.

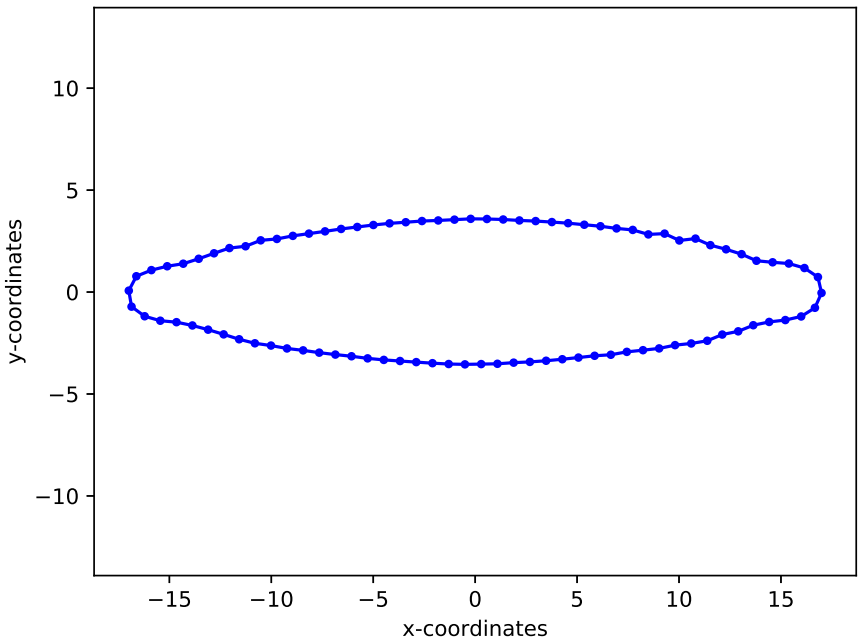


Figure 1: Plot of the first cell on the diatoms dataset

Figure 2 has all 780 cells plotted on top of each other. Each color represents a different cell shape.

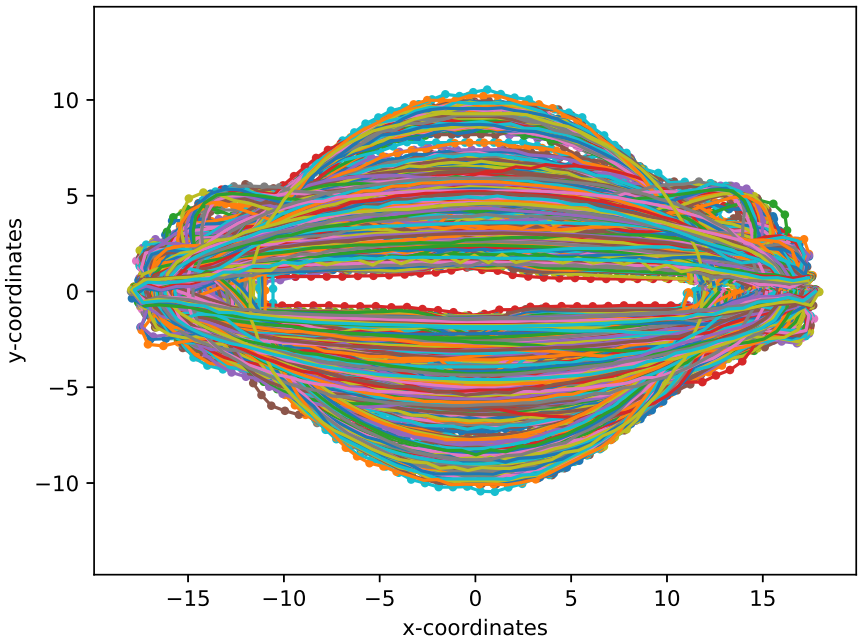


Figure 2: Plot of all 780 cells on top of each other

From the plot on Figure 2, we can see that the shapes of the cells tend to be rounder in the middle, and that they have small protuberances on the outer edges of the cell. It is also possible to see that there are many different shapes of cell, some are slimmer and some are larger.

## Exercise 2 (Visualizing variance in visual data)

Figures 3, 4 and 5 show the spatial variance of the cells of the first three principal components. The variance was done with a color map of different tones of blue to illustrate the temporal development of the cell shape.

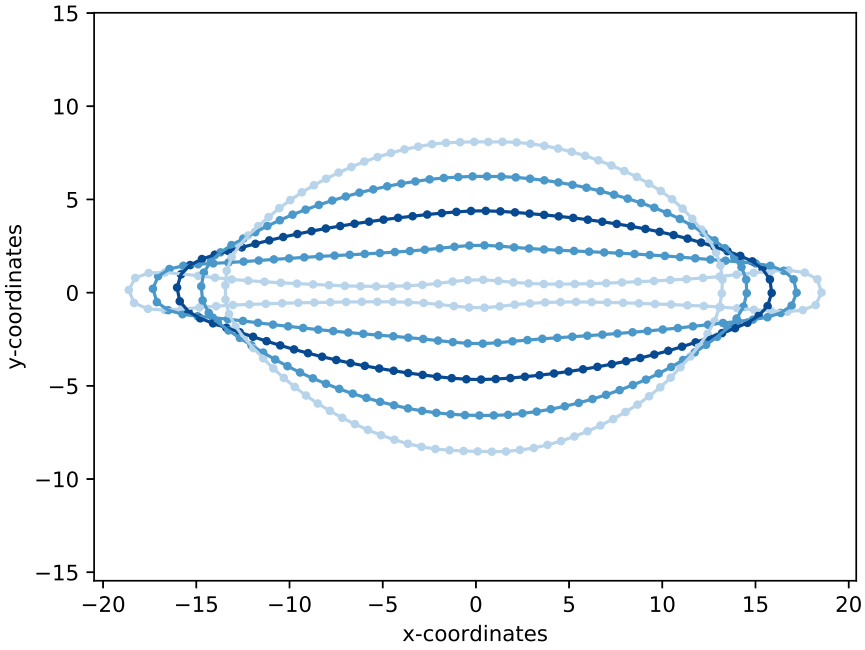


Figure 3: Spatial variance for the first PC

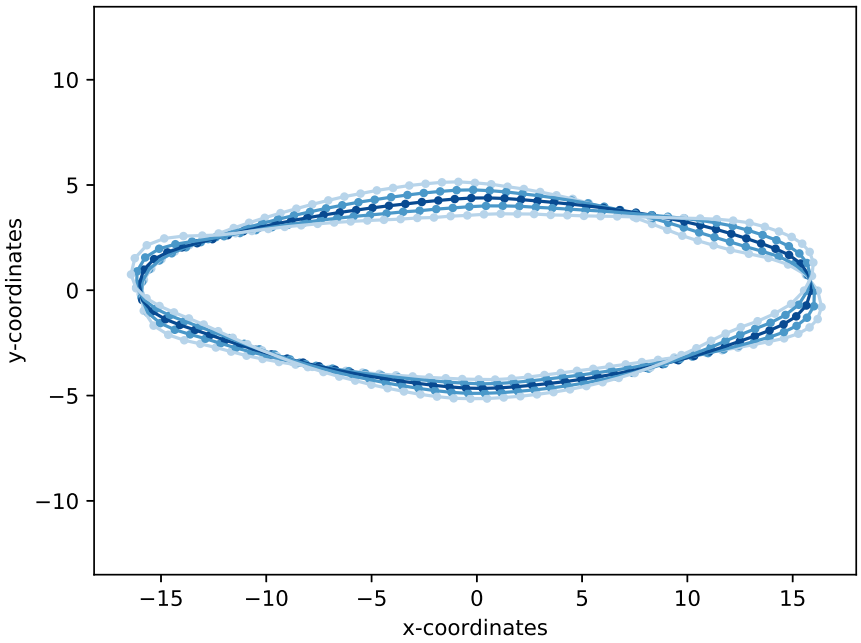


Figure 4: Spatial variance for the second PC

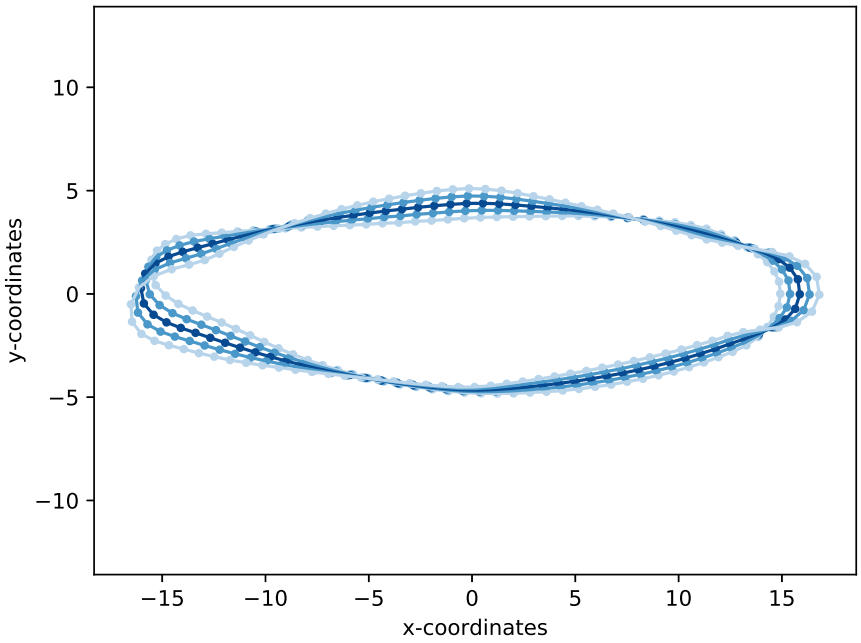


Figure 5: Spatial variance for the third PC

From visual inspection of the three figures we can see that the first principal component is the one that captures the most variance in the temporal development of the shape, since they are much slimmer or larger than the mean (the deep blue color). For the other two principal components, it seems that not a lot of variance is captured, because the shapes stay practically the same in the temporal development.

## Exercise 3a (Critical thinking)

The following answers are based on research and intuitive understanding of PCA and preprocessing of datasets. It's important to keep in mind that data pretreatment is problem dependent.

### i) Centering

Centering the data prior to PCA is not necessary. When we perform PCA we compute the covariance matrix, and that process already centers the data, so doing this preprocessing would be extra computation.

### ii) Standardization

The goal of PCA is capturing the total variance of a given set of variables as a smaller dimensional problem, and so it will require that these input variables have similar scales of measurement. Since calculating the covariance matrix only centers, but doesn't standardize the data, it may be a good idea to do it so that the largest scales don't overwhelm the PCA, since it is a variance maximizing process. Without this preprocessing, it could seem like one PC captures all the variance in the data, and that affirmation might not reflect the true behavior of the data. In conclusion, by doing so, we give the same level of importance to all variables of our dataset, so it could be a good idea for very differently scaled datasets.

### iii) Whitening

Whitening as a preprocessing of data prior to PCA is not a good idea. Whitening is a transformation that disrupts the covariance matrix, and thus we would have uncorrelated features that have all the same variance. With PCA we want to discover the principal components that capture the total variance of the dataset, and with this treatment of the data, our analysis would be disrupted. On the contrary of whitening prior to PCA, PCA can actually be helpful prior to the whitening process, as we need the eigenvector decomposition of the covariance matrix to calculate the whitening matrix.

## Exercise 3b (Critical thinking

The following figures are visualizations of the toy dataset after running PCA and projecting it onto the first 2 PCs. Figure 6 shows that projection for the whole dataset and Figure 7 shows it for the dataset without the last two data points.

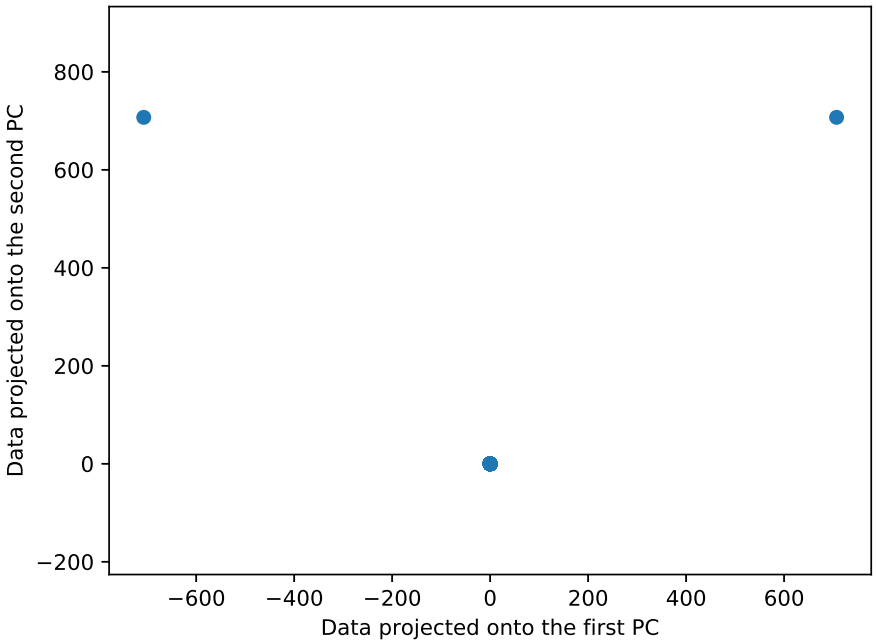


Figure 6: Toy dataset projected onto the first 2 PCs

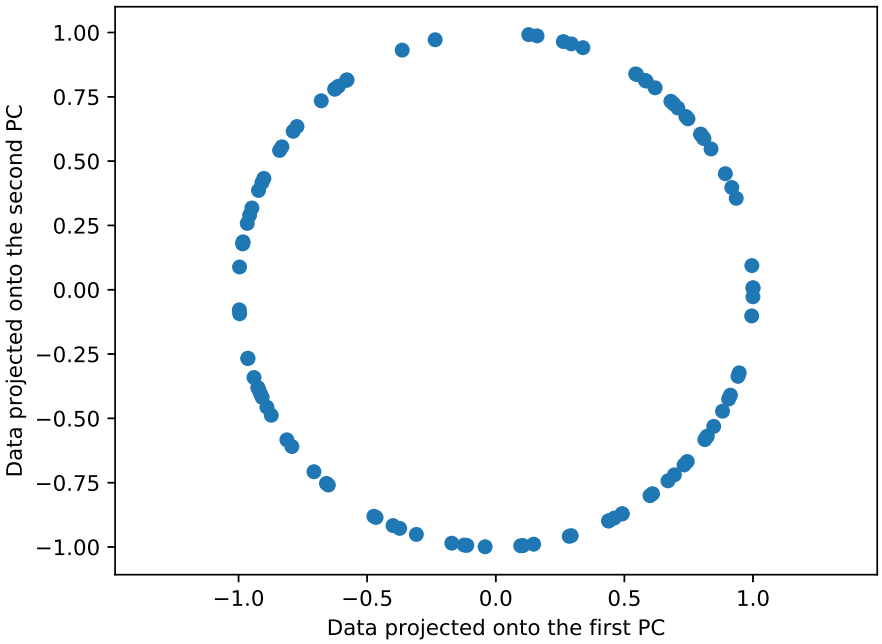


Figure 7: Toy dataset without last 2 data points projected onto the first 2 PCs

It can be seen in Figures 6 and 7 that PCA doesn't handle outliers very well. Since the last two data points on the toy dataset were very different from the rest, it overwhelmed the analysis and we don't get a good projection. As said in the exercise before, we could standardize the points, if that is what the problem demands. After removing the outliers and running the same process, the hidden structure of the dataset is revealed.

## Exercise 4 (Clustering)

For this exercise, the implementation (software used) is: first, the `IDSWeedCropTrain.csv` dataset was read and the labels were separated from the features. Then we project the XTrain matrix onto the first 2 PCs. We then plot that, with different colors for points that have different labels (0 or 1). After that, we run the `scikit-learn` K-means clustering algorithm, project and plot the cluster centers on the same plot as the projected XTrain data. Figure 8 represents that plot.

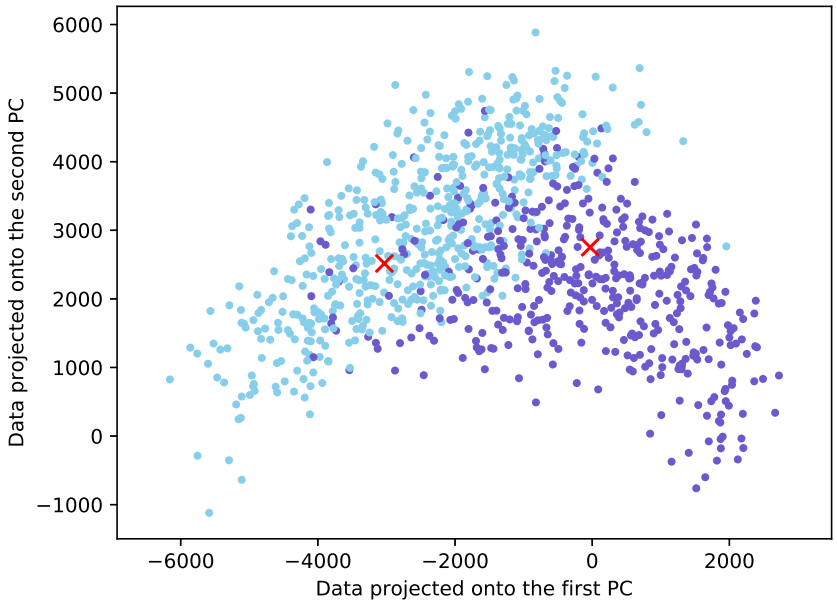


Figure 8: Crop dataset and cluster centers projected onto the first 2 PCs

The projection of the cluster centers is represented by the following matrix, where the first column is the first cluster center and the second column is the second cluster center.

$$\begin{bmatrix} -3029.74311035 & 2516.37326598 \\ -29.64633571 & 2752.75699867 \end{bmatrix}$$

Since we're dealing with a dataset that is trying to classify images of pieces of land as crops or weeds, we can see that there is still some mixing of the data points, but we do have two defined clusters that we can see in Figure 8. The clusters obtained by the K-means algorithm are definitely meaningful, and we can see that the centers are each inside an area of mostly one color.