

2 Principal Component Analysis

You can use external libraries for linear algebra operations but you are expected to write your own algorithms.

2.1 Exercise 1

Use the `Dry_Bean_Dataset.xlsx` available on the [github](#) page of the labs.

- Divide your dataset into a train and a test set.
- Preprocess the data by centering the variables and dividing them by their standard deviation.
- Write your own algorithm to perform PCA on the variables.
- Using the training set, obtain and plot the eigenvalue spectrum using the log-scale for the y-axis. What number of principal components would you select?
- Project the data (training set) in the first two principal components and color by class. Do it also for three principal components.
- For an increasing number of principal components (1 to 16):
 - Apply a multinomial logistic regression to learn a model on the training set (use `sklearn.linear_model.LogisticRegression`).
 - Transform the test set with the matrix learned from the training set. Make a prediction with the logistic model learned.
 - Assess the quality of the predictions and comment on the results.

2.2 Exercise 2

Try to apply PCA to the Swiss Roll dataset ($n = 1000$) and plot the projection on the first two principal components. Choose an appropriate color scheme for visualization and comment on your results.