Report Introduction to Data Science – Assignment 3

Exercise 1 - Performing PCA

In Figure 1 is shown the scatter plot of the data points from the murder dataset after being centred on its mean and in violet are shown the eigenvectors returned after applying the PCA algorithm.

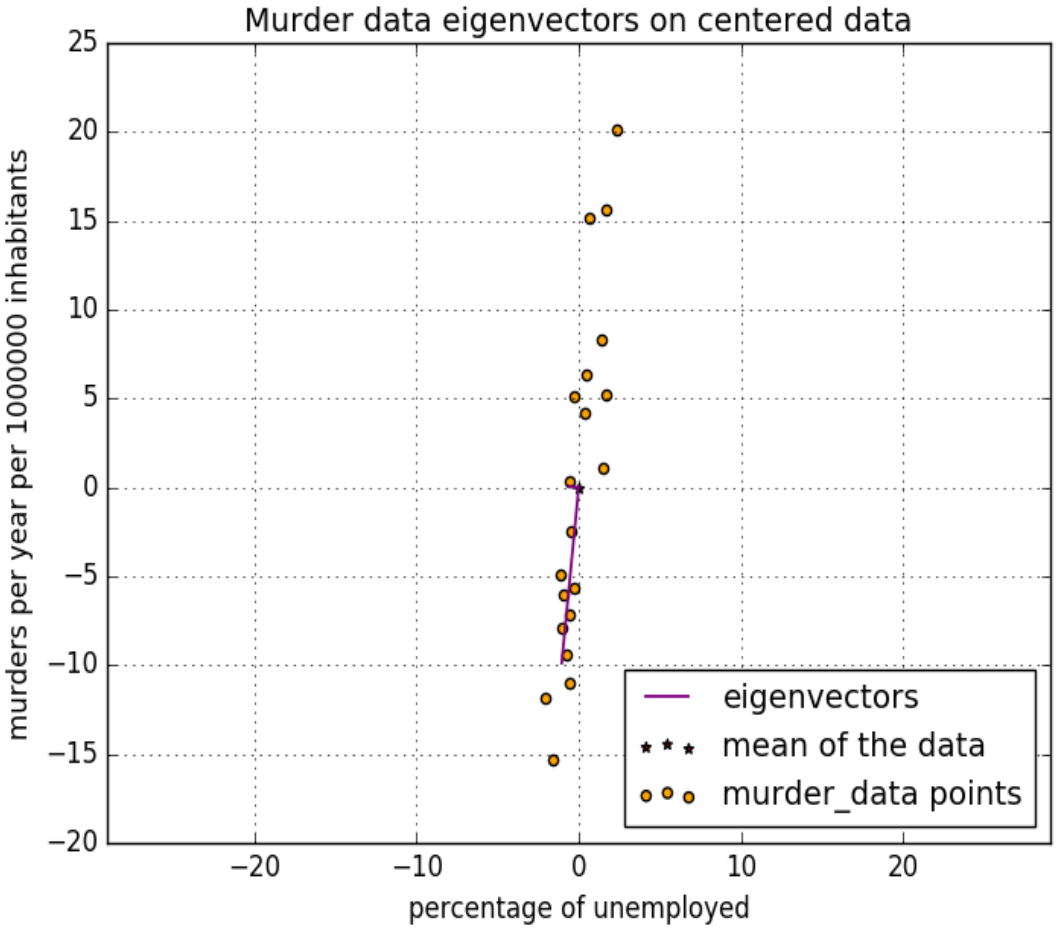


Figure 1: Plot of the eigenvectors and the data points from the murder dataset

In Figure 2 the variance from the pesticide data set is plot against the PC index. It is possible to notice that he variance stabilizes after the 5th components.

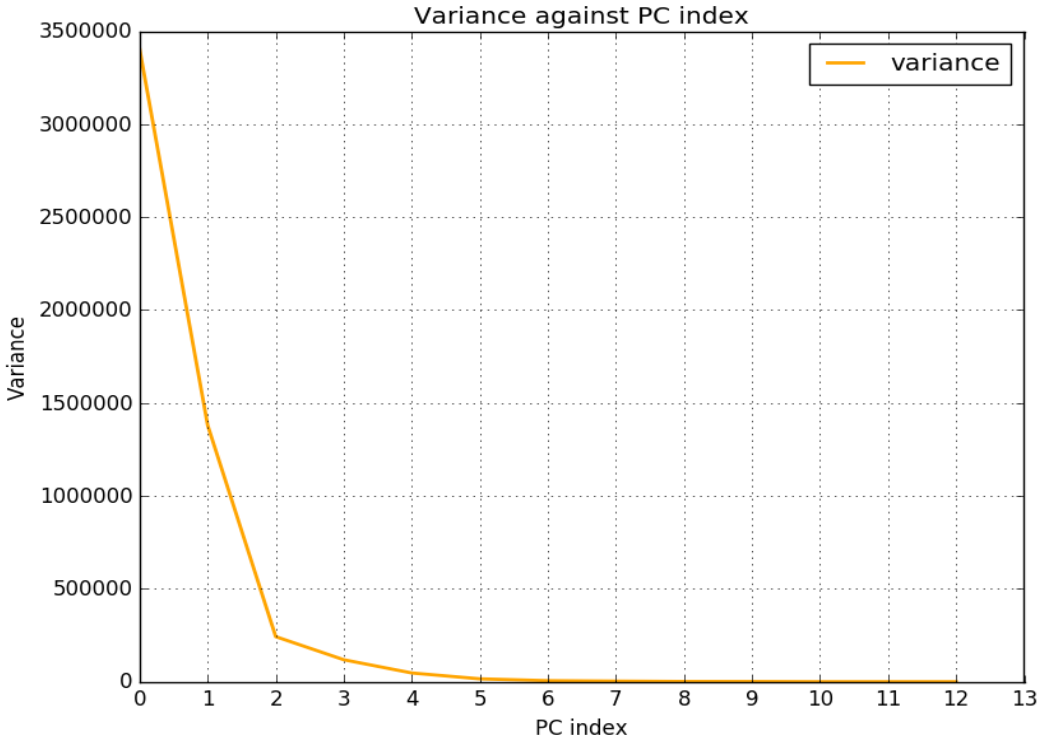


Figure 2: Variance against PC index

In Figure 3 the cumulative variance from the pesticide dataset in percent has been plotted against the PC indexes.
From the plot we can say that 90% of the variance is captured by one component; furthermore to describe 95% we need almost 2 components.

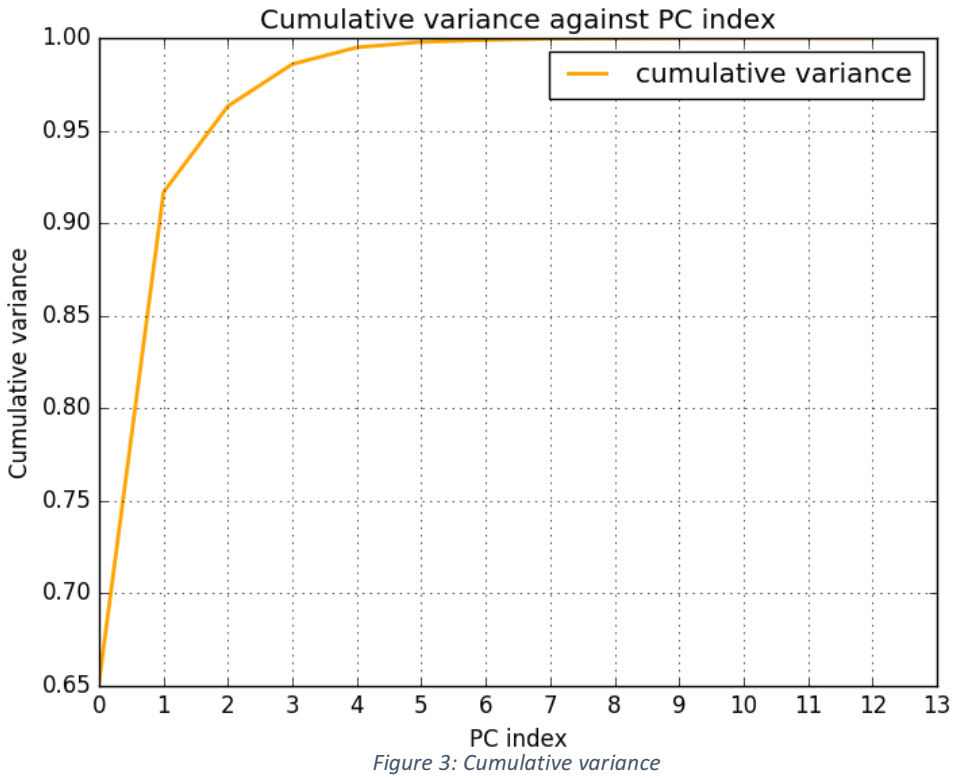
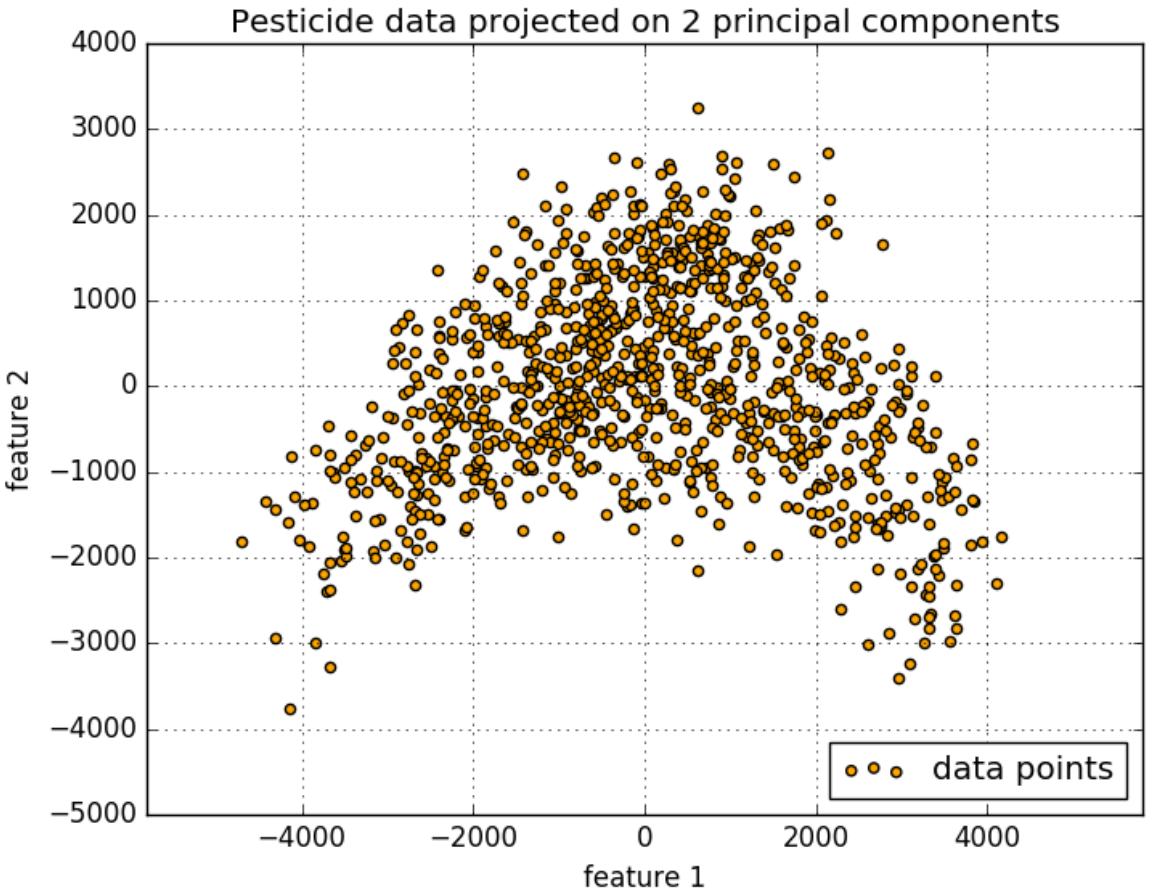


Figure 3: Cumulative variance

Exercise 2 - Visualization in 2D

The plot below shows the pesticide data projected on the first 2 principal components.



Exercise 3 – Clustering

The software used is a manual implementation of the k-means algorithm, following the guidelines from the task in the assignment.

The two cluster centres found are:

[5.70726496e+00 4.93012821e+01 7.92408120e+02 3.85595940e+03
3.38821368e+03 1.35652778e+03 2.91737179e+02 1.29989316e+02
6.86111111e+01 3.81880342e+01 1.87692308e+01 4.13461538e+00
4.42307692e-01]

And

[2.19924812e+00 1.40018797e+01 1.73727444e+02 1.40094549e+03
3.18759962e+03 2.62043985e+03 1.00147368e+03 6.31413534e+02
4.95295113e+02 2.95238722e+02 1.45689850e+02 2.91466165e+01
2.82330827e+00]