University of Crete

Msc Bioinformatics

Methods in Bioinformatics

1st Project-PCA

Name:
Sofia Kampaki

March 15, 2019

Part 1: Theoretical Problem

Question a

Eigendecomposition PCA

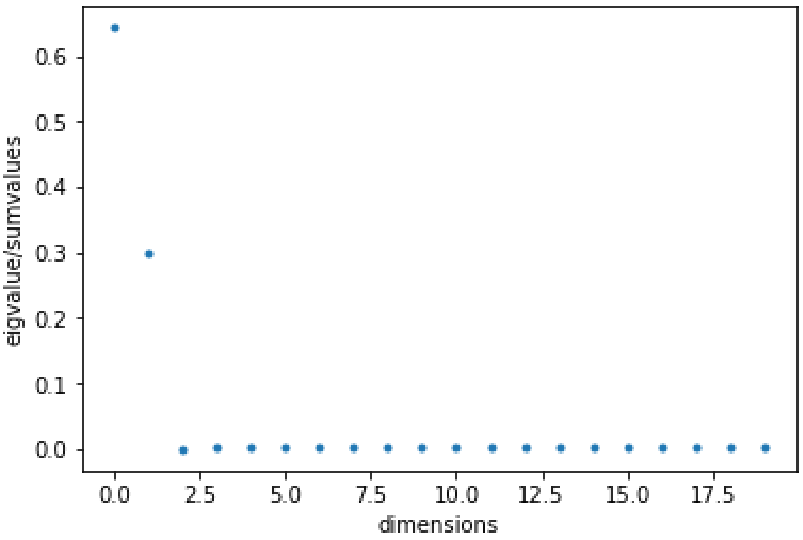

Figure 1: Eigen-decomposition pca - finding the dimensions

Here we can see that the ideal number of dimensions to keep is 2(maybe 3, but with 2 we get better results)

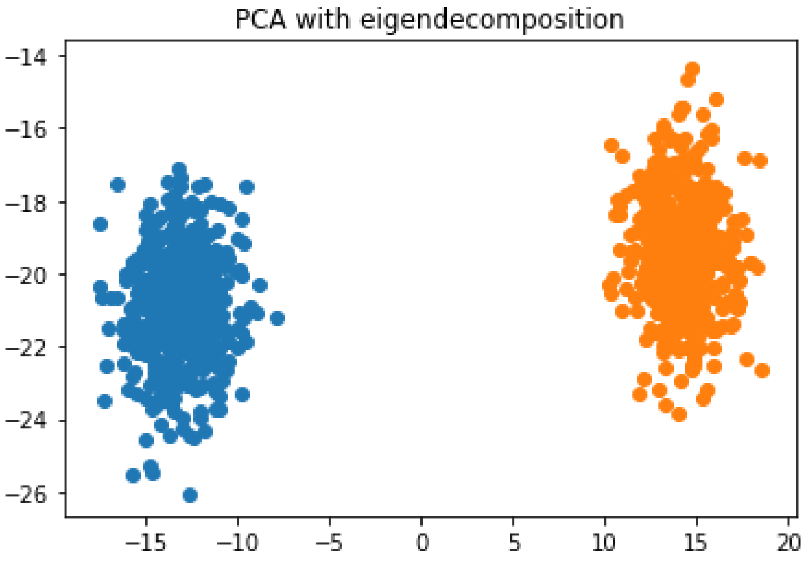

Figure 2: Eigen-decomposition pca - plotting the dimensions

Here we can see that when plotting by the first two principal components, some ”clusters” appear and it is linearized

SVD PCA

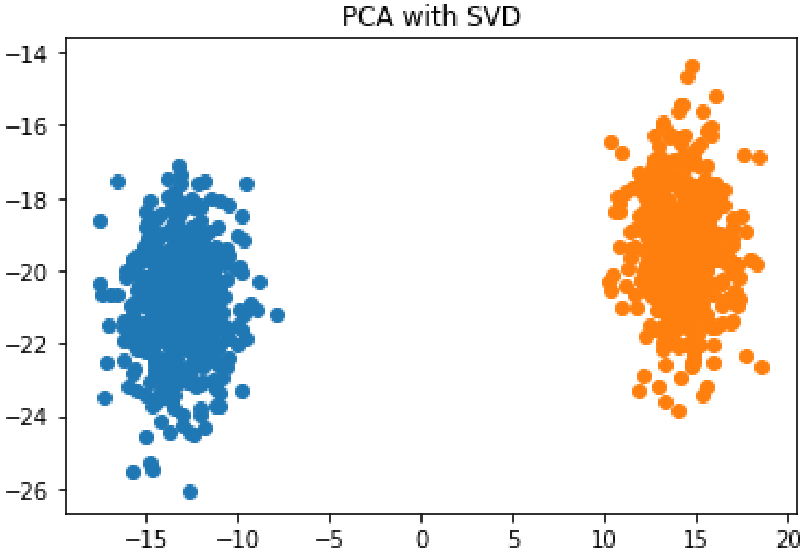

Figure 3: SVD pca - plotting the first two components

Here we can see that when plotting by the first two principal components, some ”clusters” appear and it is linearized and looking very similar to the eigendecomposition pca plot

PPCA

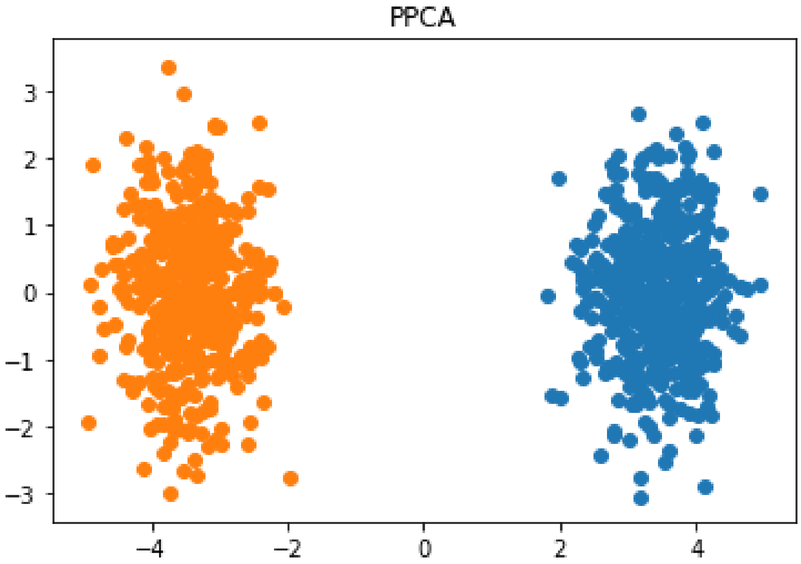

Figure 4: PPCA with sigma sguare not zero - plotting the first two components

The plot of the PPCA algorithm with sigma squared not zero has very similar results as the previous two pca algorithms. In all cases we plot the first two principal components because from the eigendecomposition method plot we observed that the first two dimensions give the most information needed and we continue plotting the first two principal components for each pca algorithm in this theoretical problem.

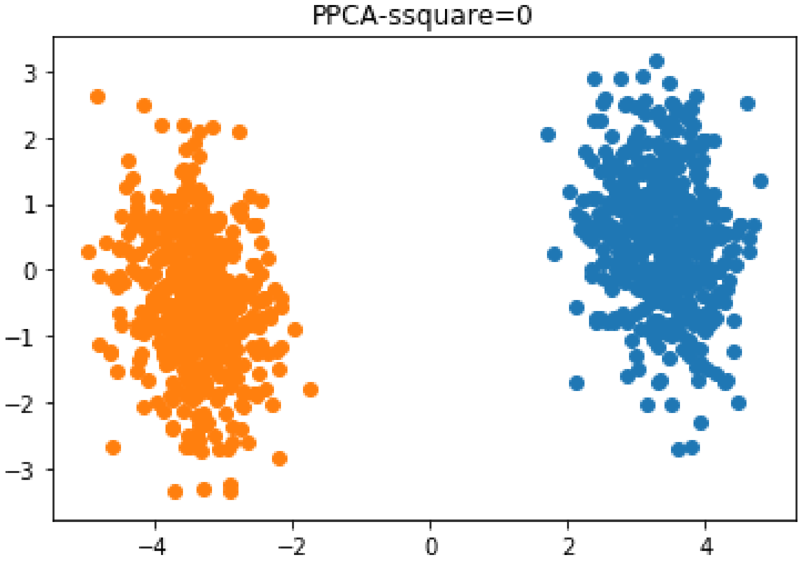

Figure 5: PPCA with zero sigma square - plotting the first two components

Part 2: Practical Problem

sklearn PCA

In this part of the project i decided to process the GDS SOFT file from the ncbi site with python commands and file processing instead of writing UNIX commands (os.system). The goal was to create a ”y” vector of mine, like the y vector that the make.blobs command produced at the theoretical part. This vector must have 84 values, each one corresponding for each person of experiment, equal to N. The X matrix that i created by cleaning the SOFT file is 84x54675 (NxD). In this experiment we want to group/label the persons according to some of their characteristics such as the gender and if they are alcoholic or not. So i create four groups, Male Control and Alcoholic and Female Control and Alcoholic, numbered from 0 to 3 in the y vector.

From the following plot i decide to keep 20 components and then i call again the PCA function to create the final X matrix (the transformed one)

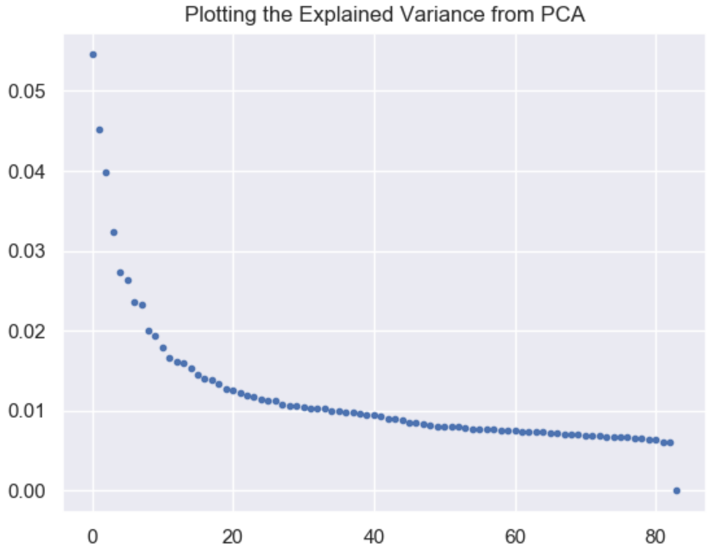

Figure 6: Plot of the explained variance ratio that the sklearn pca produced

After that, i am choosing to plot the first two components and i can see that its not very clear as the plots from the theoretical part, the points of different groups are very close to each other. Actually, that is supposed to happen because we decided to keep 20 components and plotting only the first two will not show us any valuable information. Not even a 3D plot can show us something very different for the same reason

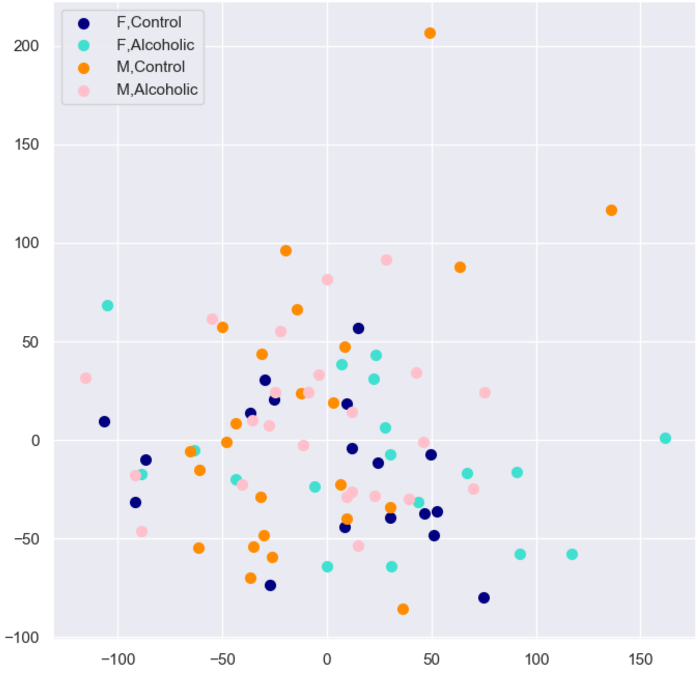

Figure 7: Plot of the first two components of the final matrix from the four groups