Report for CS215 Assignment 2

AMEEYA RANJAN SETHY - 200050006

October 12, 2021

Question No - 3

Part 1

Principal Components Analysis (PCA) is an algorithm to transform the columns of a data set into a new set of features called Principal Components. By doing this, a large chunk of the information across the full data set is effectively compressed in fewer feature columns. This enables dimensionality reduction.

And we can see that by taking very few dimensions, we can extract maximum data of the original data set. PCA actually gives different modes of variation for a given set of data and the actual variation(almost 90%) is stored in first few modes of variation.

So Now lets consider a case when random variables X and Y are almost linearly related with slight errors due to which they do not lie completely in a straight line. But by the theory of PCA if we apply PCA and take the mode of variance with the largest variance, it should represent a linear relationship between the random variables X and Y because 1st mode of variation should represent the direction of maximum variation and in this case, maximum variation is along the straight line. So the algorithm for implementing this approach is described below:

Step 1: First we import the data from the given file in variable **A** and the data set of random variable x in the variable **X** and the data set of random variable y in the variable **Y**. Both X and Y are of of size lets say $p \times 1$ where p represents the number of points in each random variable.

Step 2: Now we create a matrix **M** of size $p \times 2$ containing X and Y as the columns of M.

Step 3: We calculate the mean of matrix M and store it in mu of size 2×1.

Step 4: Now we calculate the **co-variance** matrix of **M** and store it in **C** of size 2×2.

Step 5: Now we find the **eigen vectors** and **eigen values** of co-variance matrix C and store them in V and D respectively where D is a diagonal matrix consisting of eigen values along its diagonal sorted in decreasing order.

Step 6: Now the required line depicting the linear relationship is along the direction represented by the first column of matrix V and passing through mu.

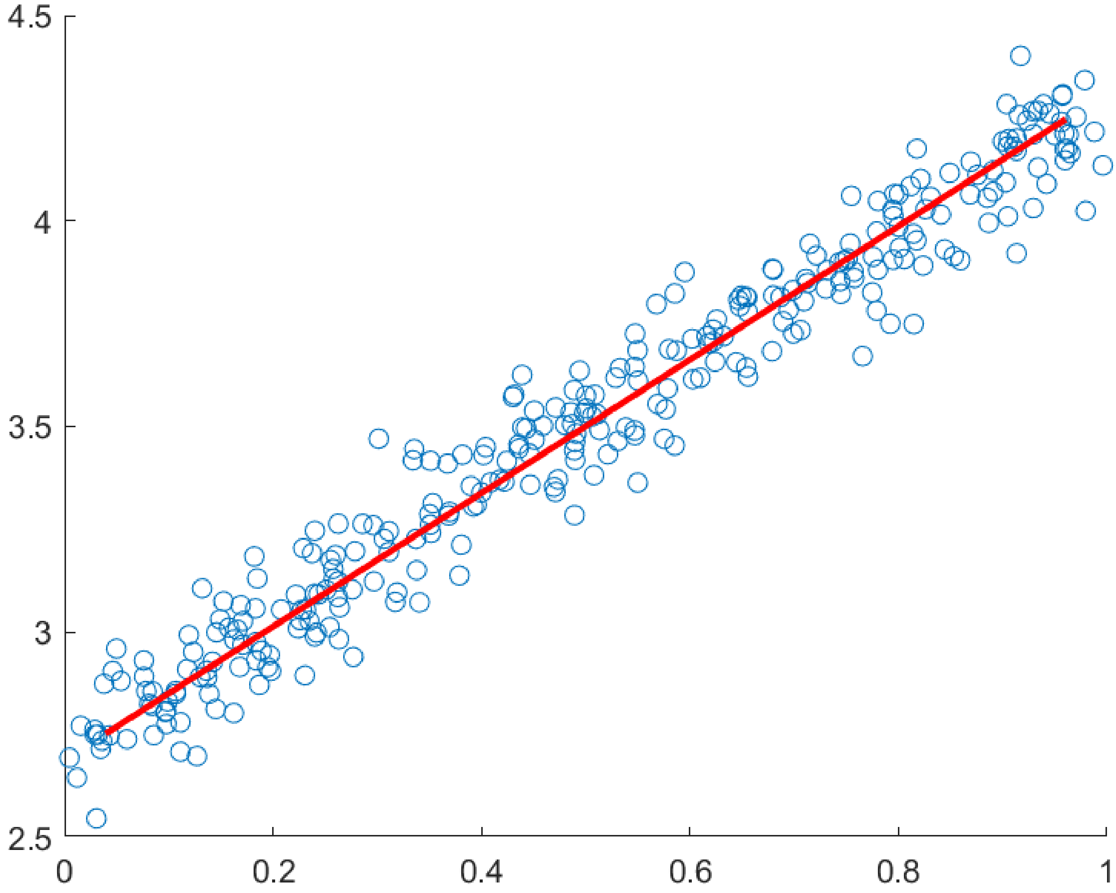
Step 7: So to plot the line, I take two points on the line as

$$P1 = \mu - 3\lambda v$$

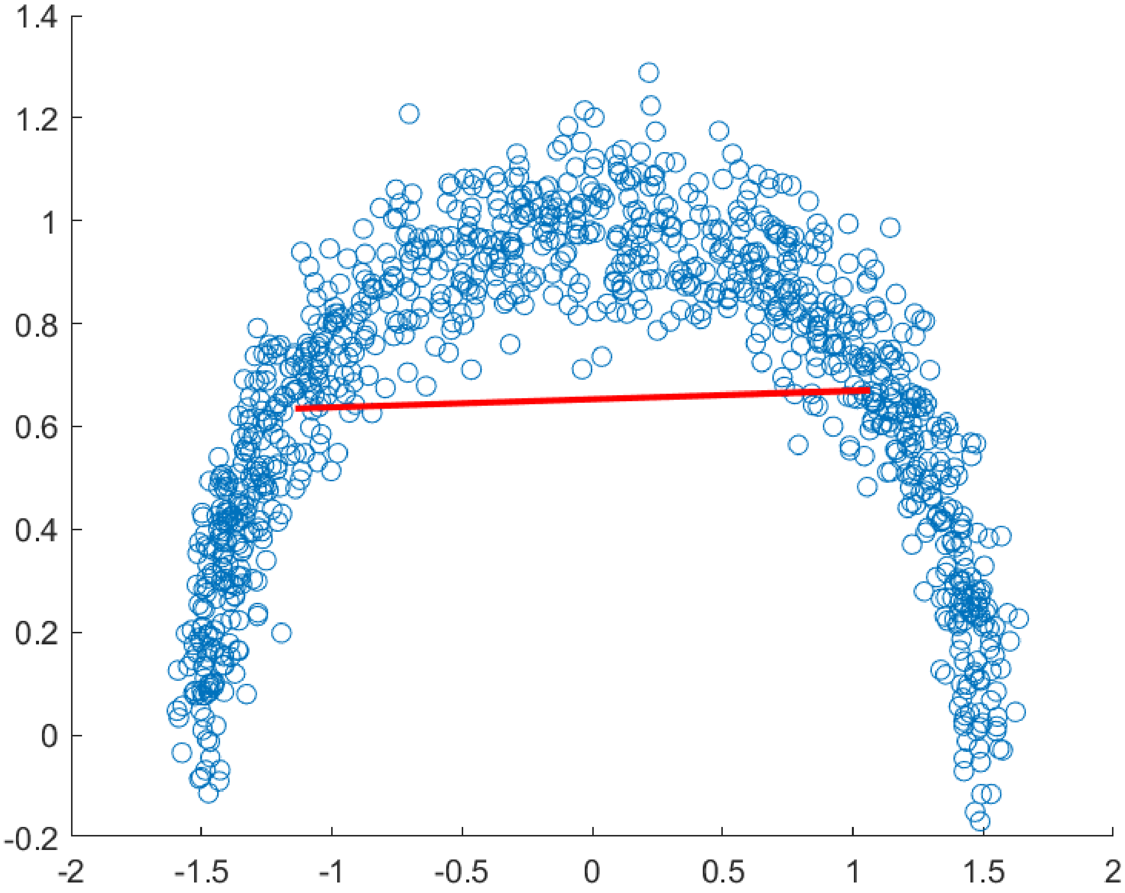
$$P2 = \mu + 3\lambda v$$

and plot the line between these two points.

First set of points



Second set of points



For the first set of data, the largest eigen value could capture about 99.14 percent of the variance of the original data but for the second set of data, the largest eigen value could capture about 91.87 percent of the variance of the original data.

So the quality of the approximation is better for the first set of data because the largest lambda could capture maximum data from the original set of data.