CS215 ASSIGNMENT 2

Utkarsh Varun

Question 3

Contents

1 Introduction

```
I and my partner had agree to use Matlab programming platform for this question.
And for plotting graphs, we use matplotlib package.

I had implement the problem statement given in assignment in 2 different files.
I had also mention some comments in the code for better understanding.

Those 2 files are as follows:-
1) q3_b.m  -> For estimating the linear relationship between random variables
  X and Y of data present in file "points2D_Set1.mat".
2) q3_c.m  -> For estimating the linear relationship between random variables
  X and Y of data present in file "points2D_Set2.mat".
```

When you run first q3_b.m and q3_c.m file, the combined plot showing scatter plot and estimated graph of line showing the relationship between X and Y

Upon running q3_b.m and q3_c.m, a combined plot of scatter plot of given data and line graph showing the linear relationship between the variable x and y of the dat file "points2D_Set1.mat" and "points2D_Set2.mat" respectively.
The plot will be saved in the file named q3_b.png and q3_c.png respectively.

I had also submit all plots in the "results" directory and all both python files in "code" directory.

2 PCA Method for Approximating Linear Relaionship between two Random Variables

Principal Component Analysis(PCA) is the process of computing the principal components of data and using them to perform a change of basis on the data.

PCA is commonly used to reduce the dimension i.e, **dimensionality reduction** by projecting the data sample into few principal components.

It can be easily shown that the principal components are **eigenvectors of the covariance matrix** of the data sample.
Thus, the principal components of data are often calculated by using the concept of eigen-decomposition of the data covariance of data.

I had also use the same concept stated above.
For 2D data smple, we will have two PCA line both **generating** from the **mean point** of data sample and of a **distance** equal to **square-root of eigenvalue** in the direction of corresponding **eigenvectors**.

For 2D, we will have (1*2) matrix mean and (2*2) matrix covariance of a 2-Dimension random variable.
So, we will get set of two eigenvalue with their corresponding eigenvector of covariance matrix.

These two sets of eigenvector will give us the direction of our **Principal Component** of 2D random variable and square root of the corresponding eigenvalue gives us the length of the **Principal Component** line.

Both of these should originated from the **mean point** of data.

3 q3_b.m Coding

I had just implement the logic stated above and plot the combined plot of scatter plot of data sample and the line graph showing the linear relationship between the given two random variables x and y in the dataset points2D_Set1.mat.

I had plotted both the two lines given in the direction by the two eigenvector of covariance matrix of length equals to square-root of the corresponding eigenvalues.

3.1 q3_b Plot

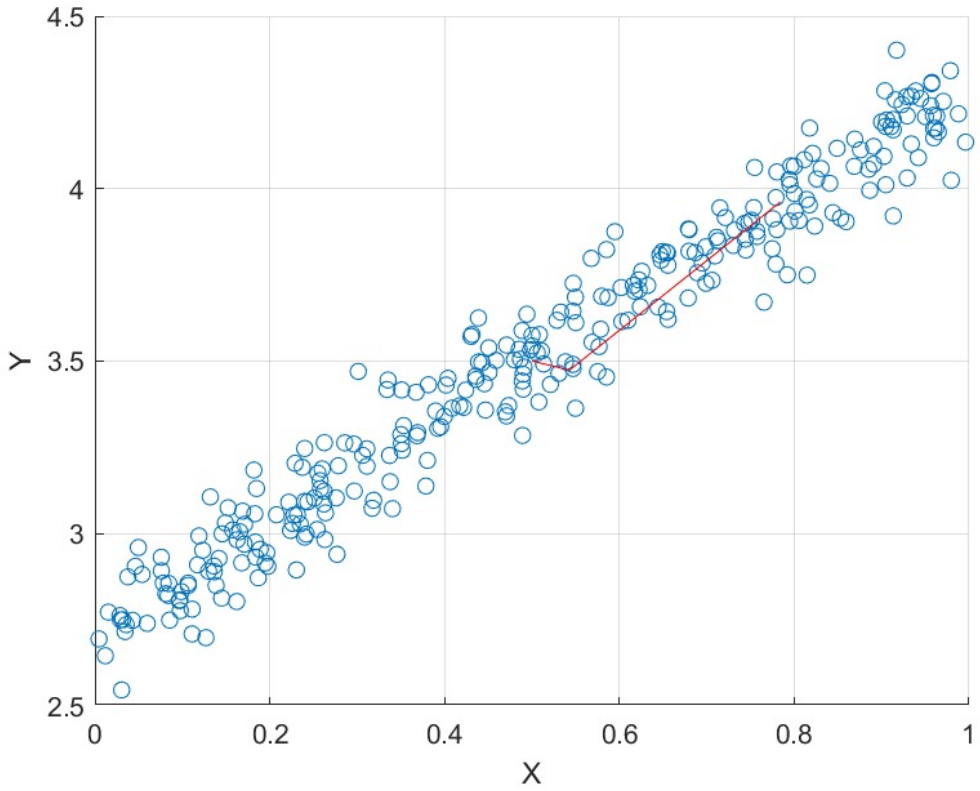

Figure 1: This is the image that i got after the analysis of dataset points2D_Set1.mat.

4 q3_c.py Coding

I had just implement the logic stated above and plot the combined plot of scatter plot of data sample and the line graph showing the linear relationship between the given two random variables x and y in the dataset points2D_Set2.mat.

I had plotted both the two lines given in the direction by the two eigenvector of covariance matrix of length equals to square-root of the corresponding eigenvalues.

4.1 q3_c Plot

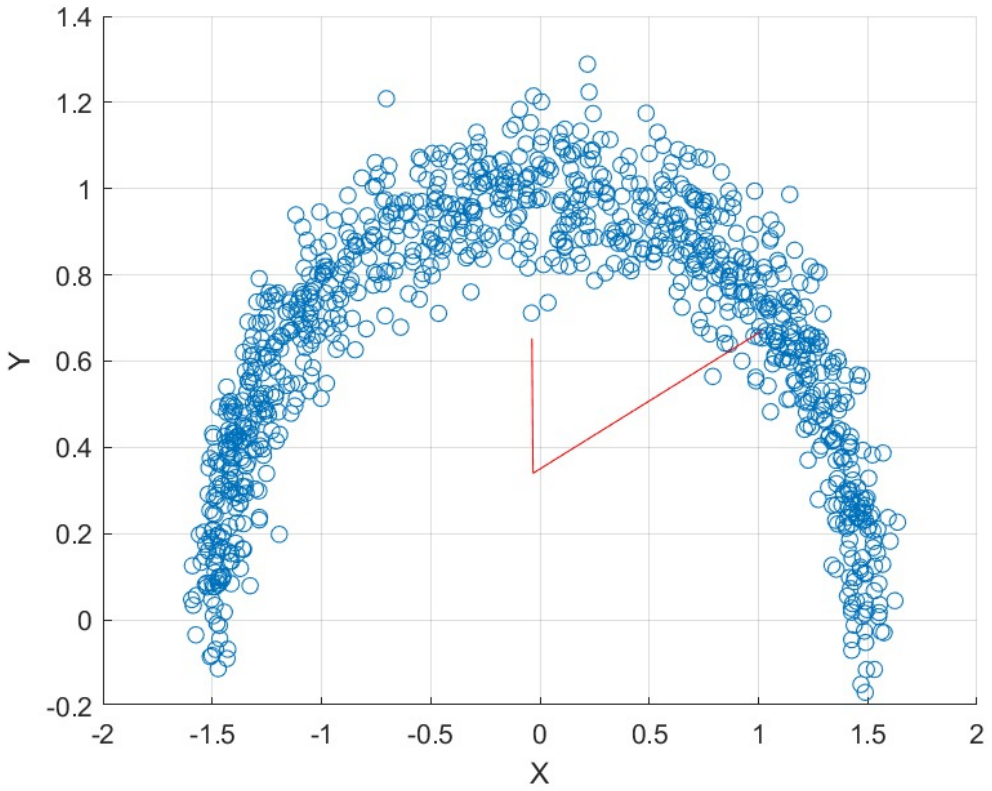

Figure 2: This is the image that i got after the analysis of dataset points2D_Set2.mat.

5 Comparison Between two above results

For part(b), the dataset has almost a linear realtionship between them(as we can see this by plottting the 2D scatter plot).
That is why the **PCA** approximation straight lines overlap on the scatter plot of dataset.

However, points2D_Set2.mat had set of data which has not the linear relationship between their 2 Random variable x and y, their exist some curve(circular) relationship.
So, when i plot the **PCA** of part(c) the **PCA** striaght lines(which are approximately showing the linear relationship) approximately were not overlapping the scatter plot of dataset.

In fact, when we reduce the dimension the **PCA** graph getting closer to the place where we can assume the whole dataset to be concentrated.
In part(c), scatter plot approximately looks like a semicircle. So, the pca lines are not overlapping normally.

We can assume that the whole dataset can be concentrated on that lines.
We can also think it like the centre of mass concept in Physics.

Thanks