Question 3

Avadhoot Jadhav - 210050027
Hrishikesh Jedhe Deshmukh - 210050073

October 2022

Q3. PCA and Hyperplane Fitting

Description

How can principal component analysis (PCA) be used to best approximate a linear relationship between random variables X and Y . Describe the method clearly, using appropriate mathematical descriptions for clarity. Your description should be clear enough to lead to a programmable implementation.

Direction of maximal variance passes through mean, So we first calculate mean of given data.

$$\mu_x = \frac{\sum x_i}{N}$$
$$\mu_y = \frac{\sum y_i}{N}$$

To calculate covariance matrix we'll find covariance of each pair as

$$Cov_{xx} = \frac{(x-\mu_x)(x-\mu_x)^T}{N}$$
$$Cov_{xy} = cov_{yx} = \frac{(x-\mu_x)(y-\mu_y)^T}{N}$$
$$Cov_{yy} = \frac{(y-\mu_y)(y-\mu_y)^T}{N}$$

So we get the Covariance matrix C as:

$$C = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

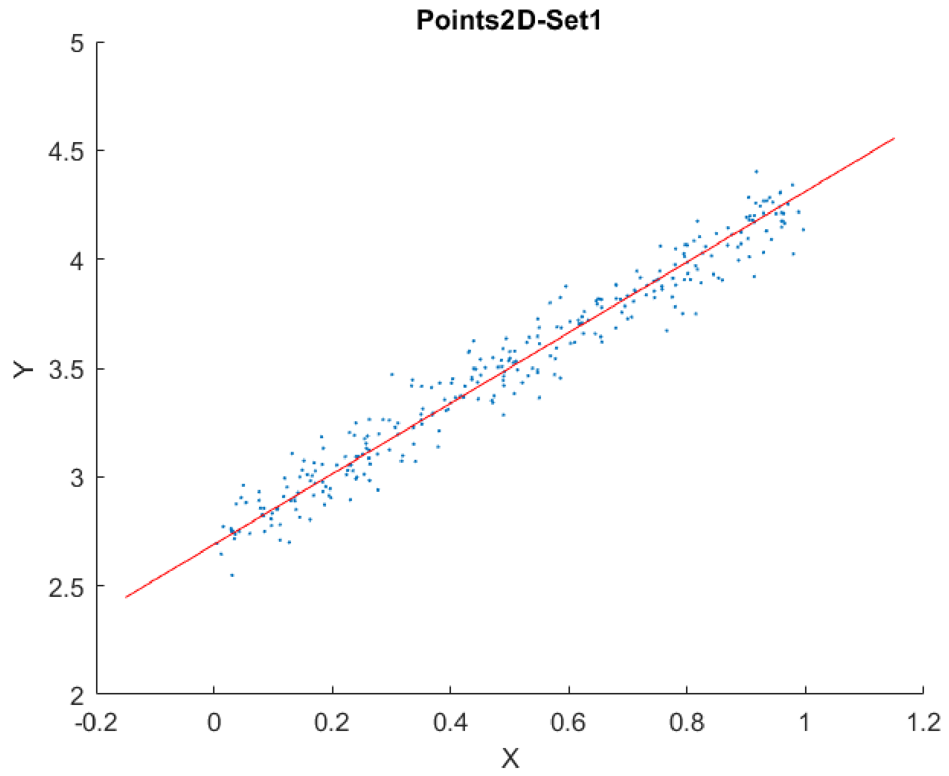
Now since C (covariance matrix) is symmetric then according to spectral theorem there exist two matrices V and D such that $C = V^{-1}DV$, Where D is a diagonal matrix with diagonal entries as eigenvalues of C and columns of V are eigenvector of C. The eigenvector corresponding to maximum eigenvalue will be the first principal component and will give direction of maximum variance. The line having slope of this eigenvector and paasing through mean will give the linear relationship between X and Y.
If eigenvector corresponding to maximum eigenvalue is $a\hat{i} + b\hat{j}$, then Line showing linear relationship between X and Y will have equation $y - \mu_y = \frac{b}{a}(x - \mu_x)$

Implementation: Using the given data we'll calculate the covariance matrix by the method given above. Now using eig() function in matlab we can get two matrices V and D, where D is diagonal matrix having eigenvalues of covariance matrix as diagonal entries and V will have corresponding eigenvectors as it's column. we'll compare both the diagonal entries and check which is larger (choosing maximum eigenvalue) and then choose corresponding

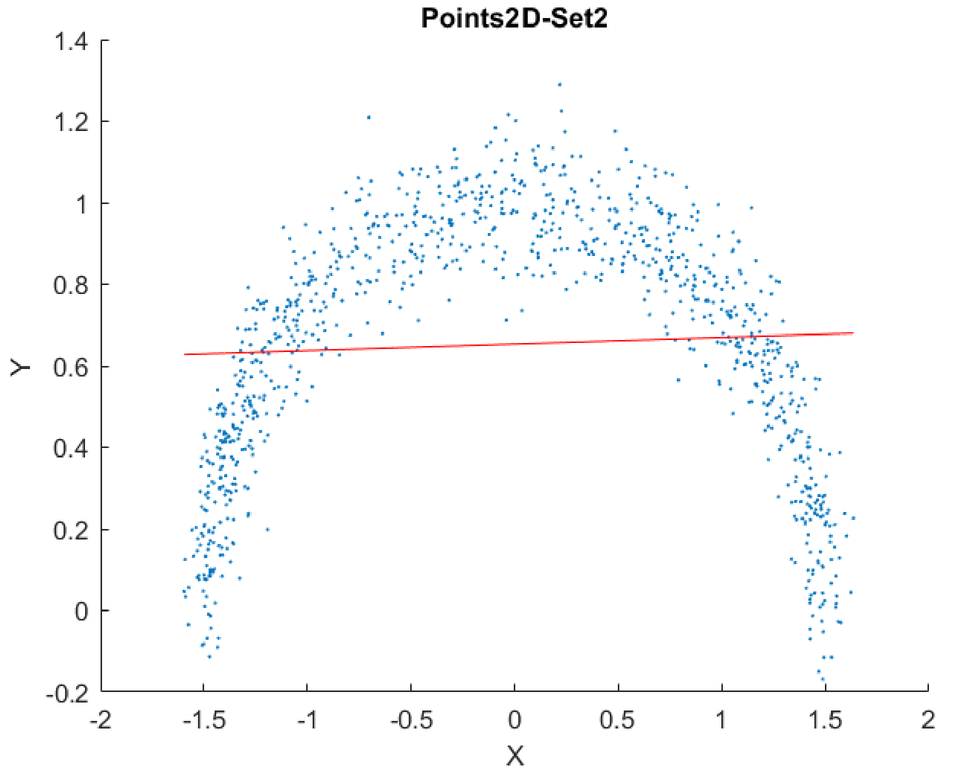
column of V. Here we got desired eigenvector.
Finally using linspace we'll generate x values in the range minimum of data to maximum of data and then calculate corresponding y by equation $y - \mu_y = \frac{b}{a}(x - \mu_x)$ and plot the line on scatter plot.

Scatter Plots

Scatter plot and linear relation between X and Y for data set1:-



Scatter plot and linear relation between X and Y for data set2:-



Comparison

Compared to the result on the other set of points, justify the quality of the approximation resulting in this question using logical arguments.

- Quality of approximation is better for dataset-1 than dataset-2

When we are approximating the relation between X and Y, we would like the points to approximately lie along a line to get the best fit. This will give us the better fit because the line that we have calculated will overlap the the points in 2D plane.

In the dataset-1 the given points approximately lie along a straight line and thus linear approximation in this case gives a good fit. While the points in dataset-2 do not lie along a line but instead lie along a horse-shoe type shape. Thus linear approximation for this case gives a very bad fit.