Question 3
PCA and Hyperplane Fitting

Atishay Jain
Gohil Megh Hiteshkumar

October 15, 2022

1 Idea

To best approximate the linear relationship between random variables X and Y using PCA.

- First calculate mean vector of the given data.
- then calculate the covariance matrix of X and Y.
- Since covariance matrix is 2 × 2 real valued symmetric matrix, it has 2 real eigenvalues and 2 real valued eigenvector according to spectral theorem.
- Now take the max value of eigenvalue and corresponding eigenvector, the eigenvector will act as vector in direction of our line and one of the point on this line will be mean.
- Now show the plot of the numbers given and the line.

2 Implementation

2.1 Calculating Mean

- for given data calculate mean of each random variable using,

$$\mu_X = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

where, x_i is data given for RV X, and N is total number of such data.

2.2 Calculating covariance matrix :

- given sample points covariance of two RV can be calculated as,

$$C = \frac{(X - \mu_x)(Y - \mu_y)}{N} \tag{2}$$

where X is column vector of sample points and Y is column vector of sample points of RV X and Y respectively. N is total sample points

2.3 Calculating Line Equation

- given a point(μ) we need a vector in direction of line to get line equation.
- To get this vector, calculate eigenvector and eigenvalues of the covariance matrix.
- Then our required vector is the eigenvector corresponding to the max eigenvalue.
- Now we know the required vector $= a\hat{i} + b\hat{j}$ then slope of line $= \frac{b}{a}$ thus line equation will be.

$$x = \frac{b}{a}.(y - \mu_y) + \mu_x \tag{3}$$

3 Observation

It can be clearly seen from the two graphs that dataset1 has betrer quality of linear approximation than dataset2

If the datapoints are bounded by a narrow rectangle then the variables are linearly related while if the data is not bounded by a wide rectangle then the variables are not linearly related since for narrow recatangle we can have a line which runs parallel to the rectangle.

4 Plots

Scatter plots of the points with overlay of the graph of a line showing the linear relationship between Y and X for both points2D_Set1.mat and points2D_Set2.mat

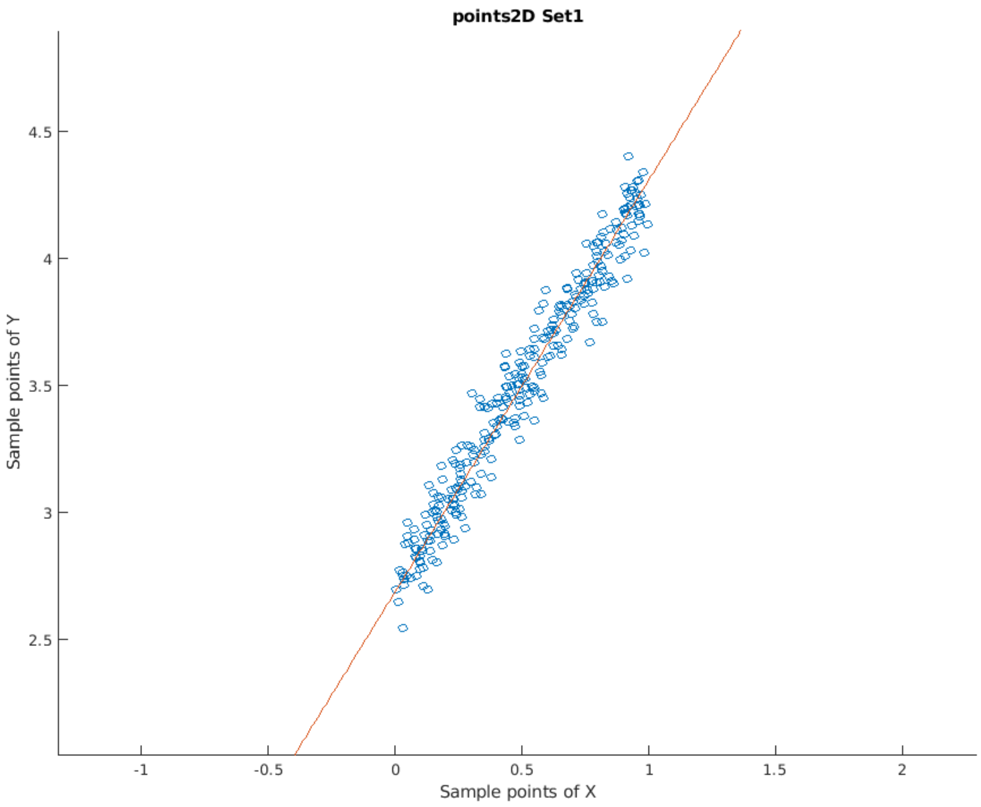

Figure 1: Scatter plot of the sample points of Points2DSet1

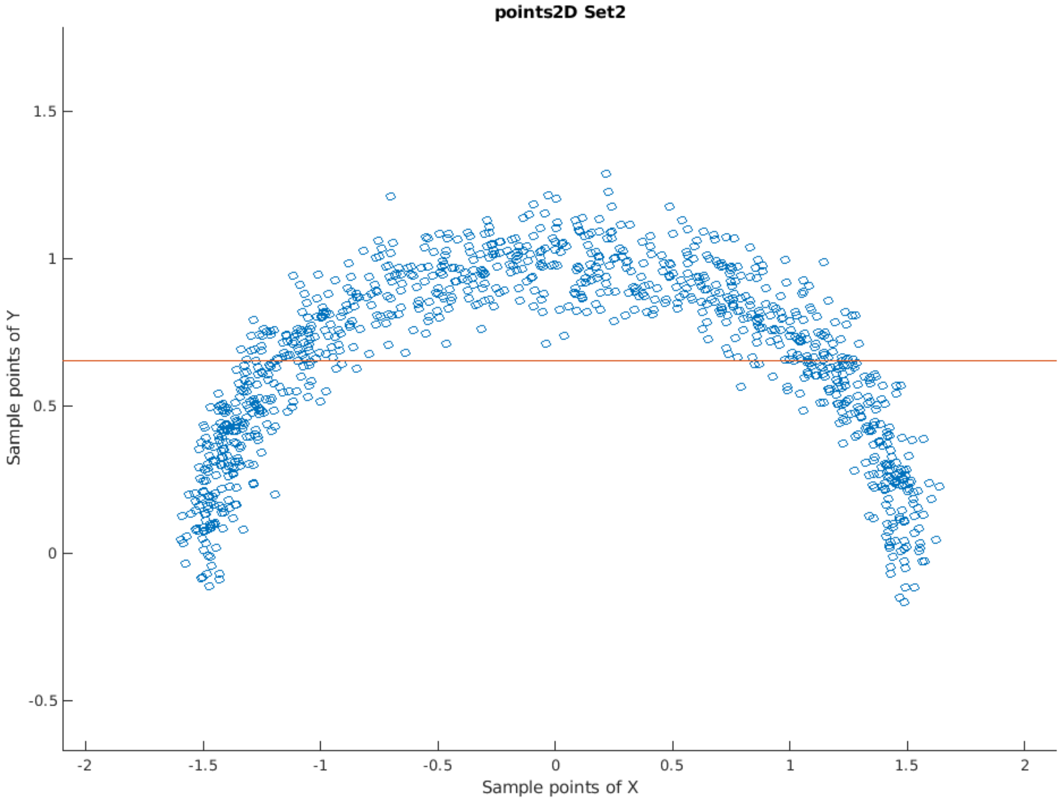

Figure 2: Scatter plot of the sample points of Points2DSet2