Question 3

3.1 : Part A

For a given 2D dataset, we need to estimate a best fit linear relationship between the two random variables from which this dataset is drawn. We will use PCA for this. Let x_i and y_i be the coordiates obtained in the i^th draw, and let $\mathbf{X}$ be a matrix of two vectos, each corresponding to the x and y coordinates.

The mean vector can be given by $\boldsymbol{\mu} = \begin{bmatrix} \frac{\sum x_i}{n} & \frac{\sum y_i}{n} \end{bmatrix}^T = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix}^T$
The Covariance matrix can be given by $\mathbf{C} = \frac{1}{r-1}(\mathbf{X} - \boldsymbol{\mu})^{\mathbf{T}}(\mathbf{X} - \boldsymbol{\mu})$
where, $r = no.\ of\ rows\ in\ \mathbf{X}$

Matrix $\mathbf{C}$ will have two eigenVectors. Let they be $\mathbf{V_1}$ and $\mathbf{V_2}$, with eigenvalues λ_1 and λ_2

The principle mode of variation will be along the eigenvector corresponding to the maximum eigenvalue. And hence this will be the most accurate approximation for a linear relation between the two Random Variables since the variation is maximum along the direction of the principle mode of variation. This also implies that the squared sum of distances of all data points will be minimum from a line passing through the sample mean, along this direction. Hence this will be the best fit line

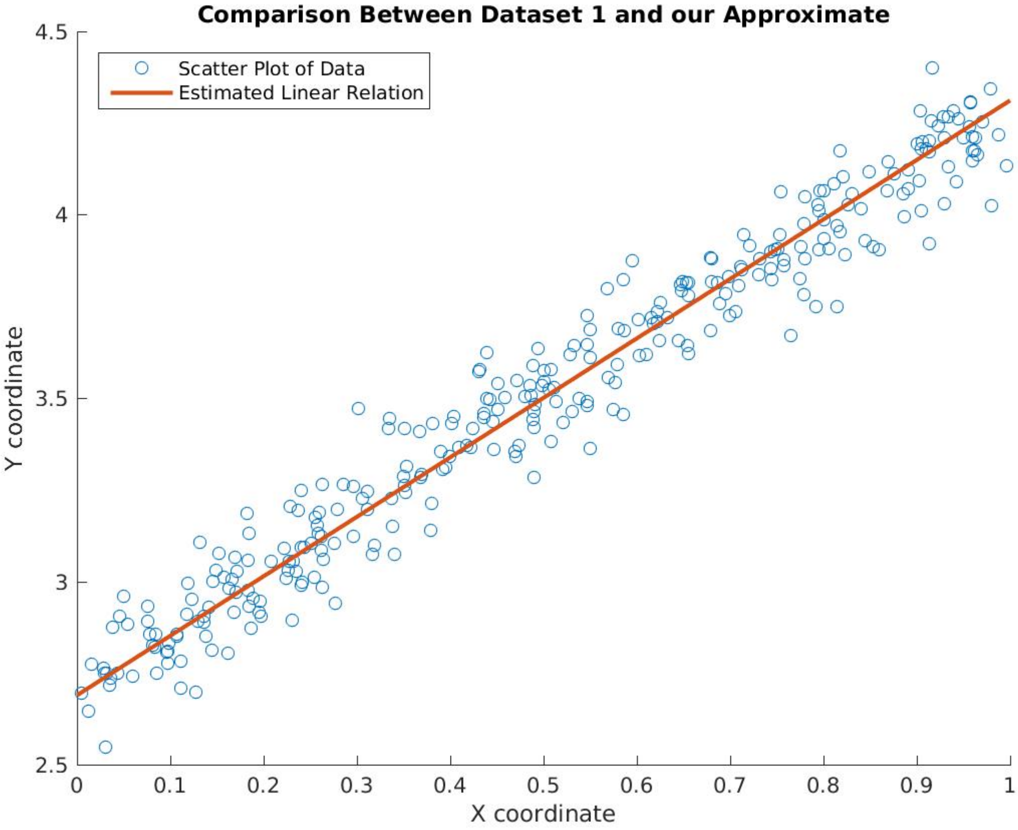
Hence, to derive the equation of the best fit line:
WLOG, let $\lambda_1 > \lambda_2$. Hence our principle mode of variation will be along the eigenvector $\mathbf{V_1} = \begin{bmatrix} v_x^1 & v_y^1 \end{bmatrix}^T$ (let)

$$\frac{y - \mu_y}{x - \mu_x} = \frac{v_y^1}{v_x^1}$$
$$y - \mu_y = \frac{v_y^1}{v_x^1}.(x - \mu_x)$$
$$y = \mu_y + \frac{v_y^1}{v_x^1}.(x - \mu_x)$$

Thus we have derived an equation for the best fit plot using PCA!.

3.2 : Part B

The plots for the first dataset along with a best fit plot showing the linear relationship between Y and X follows:



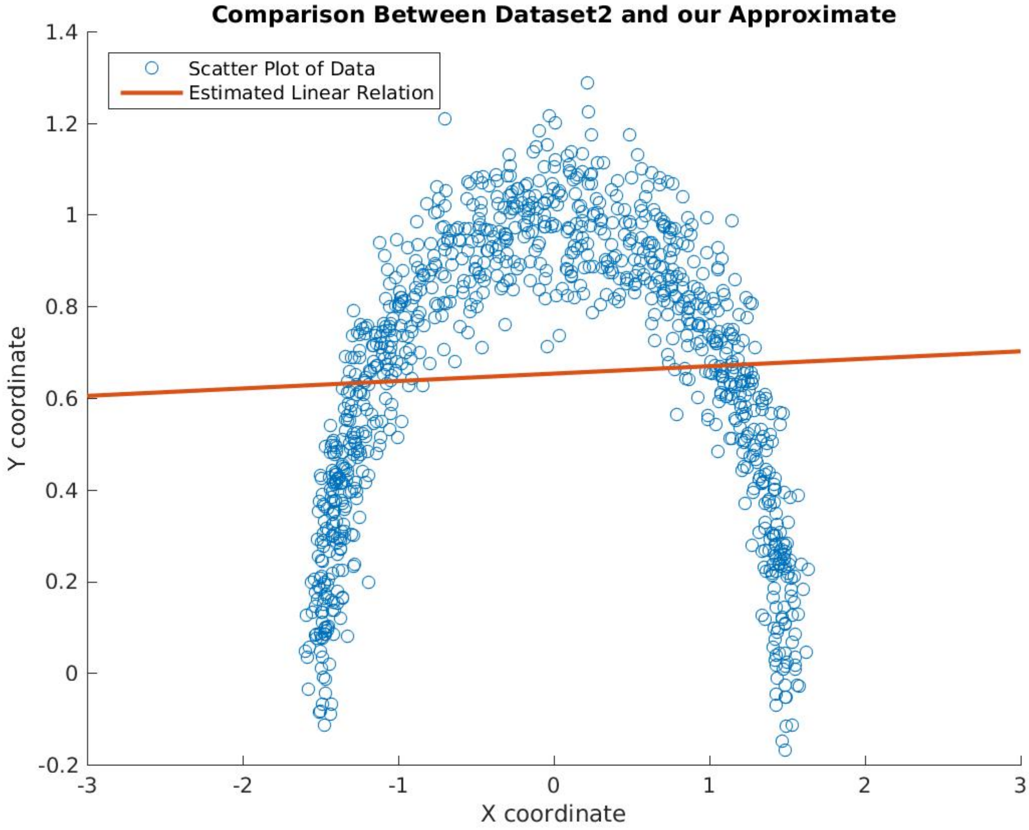
3.3 : Part C

The same procedure was repeated for another Dataset.

But in this case, the results were not very good. The second dataset was drawn from Random Variables where X, Y coordinates did not have a linear relationship between them. Hence fitting a straight line into this data is obviously a bad idea.

Upon performing PCA, we got a best fit 'straight line' through the data but this is not a good model. Instead of a linear model, perhaps a quadratic model could have given a better result.

The plots for the second dataset along with a best fit plot showing the linear relationship between Y and X follow: (PTO)



3.4 : Usage of Code

The following are the instructions for the usage of the code:

- Load the code present in 'submission/code/q3/q3.m ' .
- In the same directory are functions implemented like myMean, myCov which return the mean and covariance of appropriate matrices. Function myeigs will return the eigenvalues and corresponding eigenvectors in sorted order
- Simply run the code in 'q3.m ' and this wil automatically create the required plots.
- Lines 22, 45 (commented by default) hava a code to save jpg files of the respective plots. Comment/Uncomment these lines appropriately according to need.