PROBLEM 3

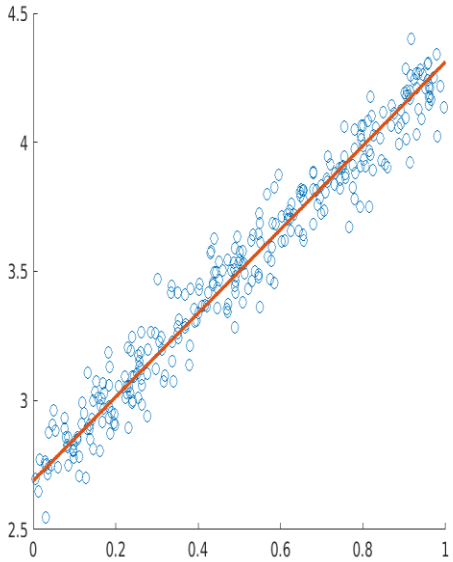
PCA is used to diminish the dimensions of a given data set. We are given (x,y) in two dimensions. Let X and Y be related as $Y = mX + c$. So we apply PCA to this dataset to reduce it to one dimension and get best values of m, c.

First we take the data set X and calculate its mean and covariance matrix using

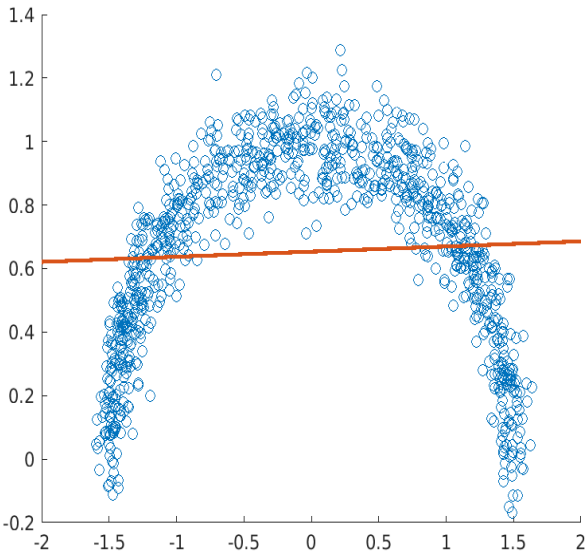
$$u = \text{mean} = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad C = \text{Cov} = \frac{1}{n}\sum_{i=1}^{n}(X_i - u)(X_i - u)^T$$

Then we calculate the eigenvalues and their corresponding eigenvectors. The eigenvector with the largest eigen value corresponds to the principal direction.

Using that eigenvector for direction and u for point, we calculate m, c.



DATASET 1



DATASET 2

Both sets of data appear to be viable candidates for PCA as they look dependent from the scatter-plot. However, only dataset 1 appears to have a linear relation whearas dataset 2 appears to be non-linear. Because of this, we see that the PCA dimension reduction to a straight line matches dataset 1 but the line does not represent dataset 2. Hence, linear relation is a good approximation for dataset 1 while bad for dataset 2.

We say that the PCA dimension reduction is "accurate" when the principle eigenvalue(s) are much larger than other eigenvalues as eigenvalues represents variance along the corresponding eigenvector. From the data given, this condition of accuracy is matched for dataset 1, however the eigenvalues of dataset 2 are not close for linear PCA. Hence, linear relation is not a good approx. for dataset 2.