CS215 Assignment2 Problem 3

Parth Pujari and Anish Kulkarni

October 2022

1 PCA to approximate a linear relationship

We are given samples of two random variables X and Y
We need to find the best linear relation approximation between them.

Covariance Matrix

First compute the Covariance matrix of X, Y which is given by

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

Here $Cov(A, B)$ can be calculated empirically as

$$\frac{\sum(A_i - \overline{A})(B_i - \overline{B})}{N-1}$$

Eigen Vector

Next step is to get eigen vectors of covariance matrix C.
As C is symmetric it has a orthonormal basis of eigenvectors.
One can use the in-built functions for this task or another way is to find roots of $det(C - \lambda I) = 0$ to get eigen values and then compute the vectors by solving linear equations corresponding to each eigen value.
Now we find the eigen vector v with the largest eigenvalue.

Line of best fit

Now this eigen vector v denotes the direction of line of best fit.
Find slope of this eigen vector $(\frac{v(2)}{v(1)})$ and call it m.
Now use the fact that our line must pass through mean values to get intercept $c = \overline{y} - m\overline{x}$.
Finally

$$y = mx + c$$

is the required line

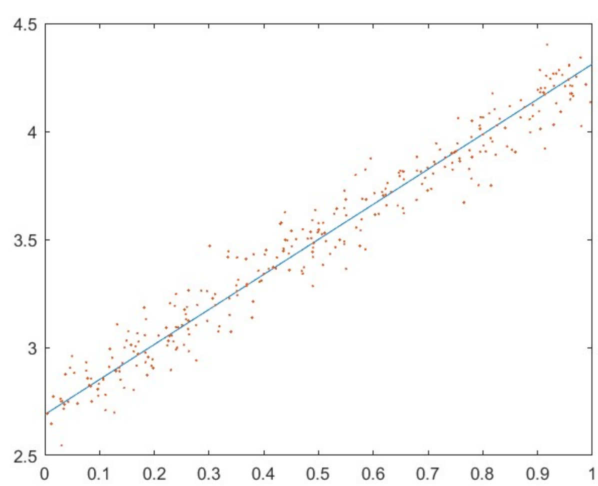

Figure 1: Set1 Points

2 Comparison

The first set clearly gave a much better result as the data was present in a linear fashion
In the second set the line doesn't match the data well because the data given was not in a linear fashion.
In precise language $|corr(X, Y)|$ was close to 1 in first case while it was close to 0 in second case.
Linear approximation of a data can accurately represent the data only when $|Corr(X, Y)|$ is nearly 1.

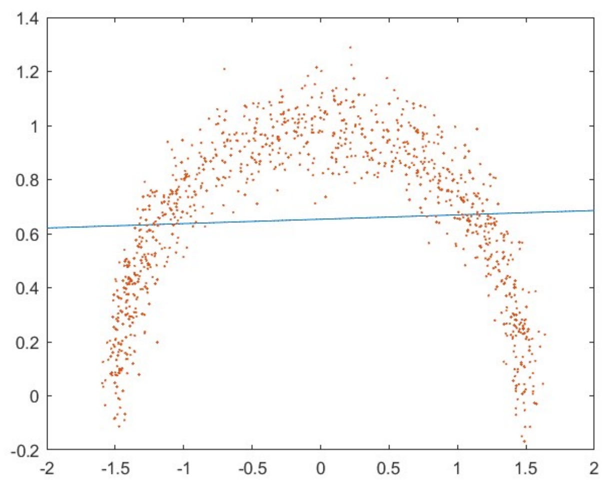

Figure 2: Set2 Points