Assignment-2 REPORT

200050040–200050075

QUESTION-3

Contents

PCA and Hyperplane Fitting

1 Linear Relationship between X and Y

We are given a Sample data of random draws drawn independently from the PDF(X,Y) which is not known.Our ain is to find an approximate linear relationship between **X** and **Y** using PCA.

So we need to find a line passing through the sample mean μ so that after **projecting** all the points onto that line, the variance is maximum.i.e, We need to find the direction of **Maximal Variance**

So if our data is $\{x_i \ \in \mathbb{R}^D\}_{i=1}^N$ and our required 'direction' is **v**(unit vector), Lets assume that we shift to a frame where the mean $\mu = 0$.So the projections of the points are $\langle x_i, v\rangle v$. And the variance is $\sum_{i=}^N \frac{\langle x_i,v\rangle^2}{N}$.

So for the variance to be maximum, we need $\sum_{i=}^N \langle x_i, v\rangle^2 = (x_i^T v)^2$ to be maximum.We can find that it will be maximised if **v** is along the **eigen vector** with maximum eigen value. It is called as the **Primary mode of variation** and the variance along that direction is given by the corresponding eigen value.

So to find that **v**

- First we need to find the sample mean and Covariance of the given PointSet similiar to Q2. Mean μ is just the average of the whole data
- If D = **2xN** matrix of the obtained 'N' data points (x,y) , then we know that the Covariance is $\mathbf{C}_{2\times2} = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$, $= (\mathbf{D} - \mu)((\mathbf{D} - \mu)^{\mathbf{T}}/\mathrm{N};$
- Then we find the 2 eigen vectors of C and their eig values. Our required 'direction' **v** would be the eigen vector with highest eigen value.
- So the required line passes through μ along the direction **v**.
- This line gives us the best approximate Linear Relation between **Y** and **X**.

2 Scatter plots and Observations

In the code, I just followed the same procedure for both the data sets PointSet-1 and PointSet-2.

- First stored the data in a 2xN matrix and found the mean and Covariance
- Then found the eigen values and picked the Primary eigen value
- Plotted the scatter plot of the data, and the line depicting the linear relation between **Y and X**

In the plots, I also marked a point at the mean of the data and also the variance(= the primary5 eigen value λ) along the Pricipal Axis shown in the solid red line. i.e, I marked a solid red line along the principal direction which goes from μ to a length = $\sqrt{\lambda}$ along the eigen vector **v**

The obtained plots are:

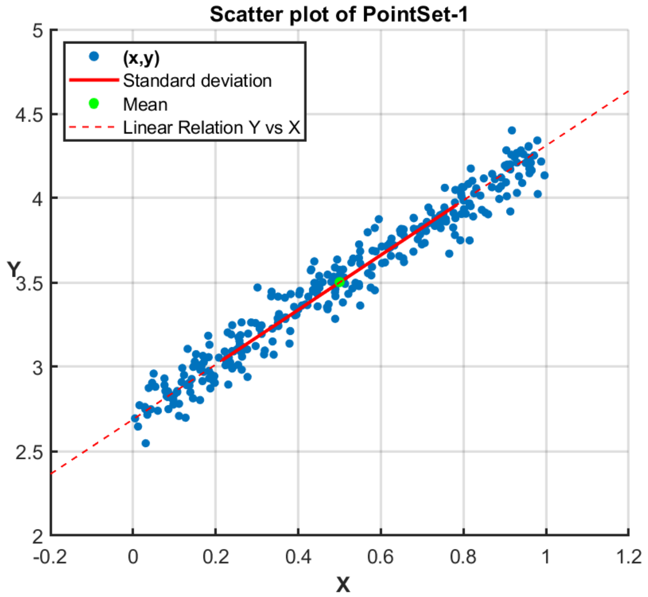

Figure 1: PointSet-1

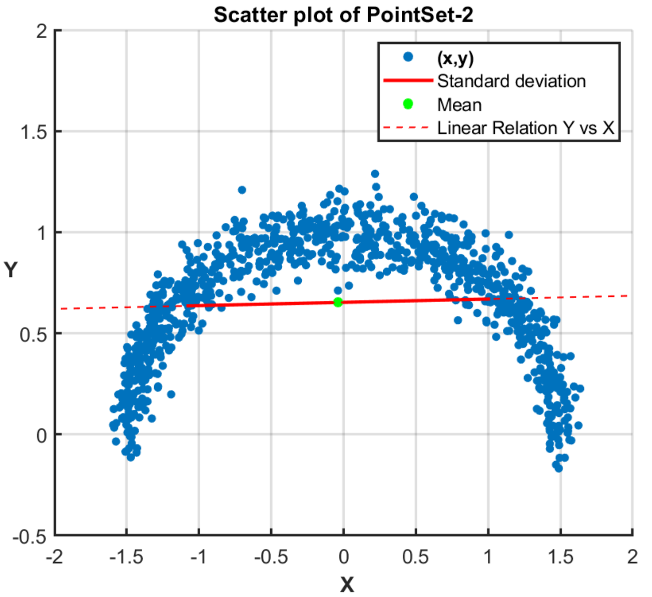

Figure 2: PointSet-2

Also the obtained line equations are:(Displayed on the console while running the code)

PointSet-1: $\mu \approx (0.5,3.5)$ and $\mathbf{v} \approx (0.52,0.85)$

So the line equation is (y-3.5)0.52 = (x-0.5)0.85 i.e, y = 3.5 + (x-0.5)(0.85/0.52),So finally the realtion between X and Y is Y = 1.63X + 2.68

PointSet-2: $\mu \approx (-0.03,0.65)$ and $\mathbf{v} \approx (-1,-0.02)$

So the approximate relation is finally Y = 0.02X + 0.65

2.1 Observations

We can clearly see that the line fits the data reasonably well in Set-1.And it captures most of the variation in the data.

But in the Set-2, the linear relation approximation is bad.We can see that the Given data is Nonlinear and so we can't have a good linear fit to the data.That's why compared to Set-1, the quality of approximation is bad here.

Also we can see that in case Set-1, the 2nd eigen value is very small compared to the 1st, Hence the Variation along other mode of variation is Very less and consequently The primary mode of variation fits the data well.

But in Set-2, the 2nd eigen value is also significant and so there is some significant variation along the perpendicular direction as well. Hence the approximation is bad in this case.

3 Code Running Instructions

`Run Q3_hyper.m file in 'code' folder to generate the set1.png, set2.png corresponding to PointSet-1, PointSet-2 respectively .` You can also find them in the 'results' folder.

I have kept the PointSet1.mat and PointSet2.mat in the 'code' folder for the program to read those data files while executing.

The program will display the mean and the direction **v** also the 2 eigen values of both the data sets on console while executing.