3.

a. $(x, y) \in \mathbb{R}^2$

$P(x, y) \Rightarrow$ Joint PDF of x, y.

$\Rightarrow$ There can be two different methods to tackle this problem.

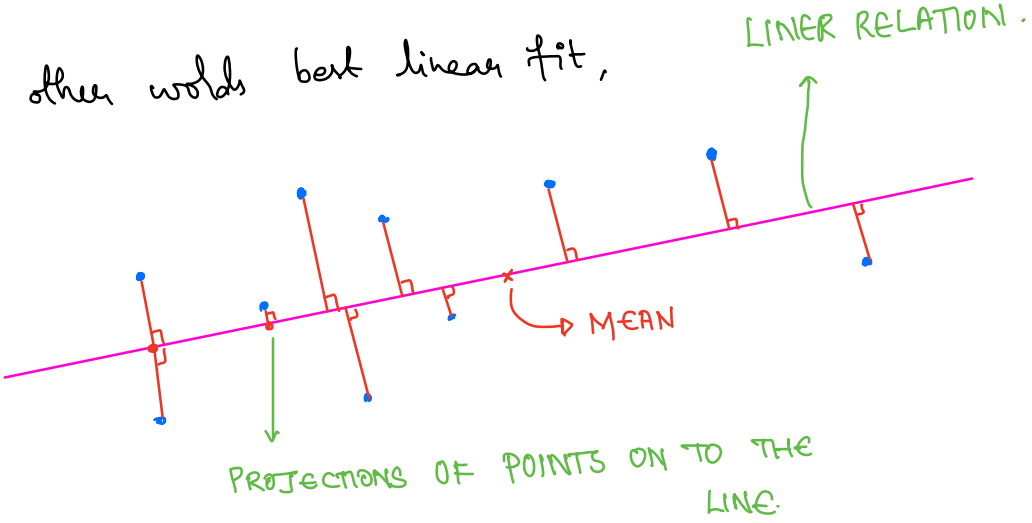
I. Using eigenvalues and eigenvectors :-

Let us say we find an eigenvector $V_{\lambda_{max}}$ and corresponding max. eigen-value λ_{max}.

We argue that this will give the line passing through at having $V_{\lambda_{max}}$ as direction.

Line with LEAST SQUARED ERROR.

In other words best linear fit,



- First we show that line passes through mean
- And then we will show that PCA component is the best direction.

Let us manipulate the sum of squared distances first:

$$e = \sum \frac{|mx+c-y|^2}{m^2+1} = \frac{m^2\Sigma x^2 + \Sigma y^2 + nc^2 + 2mc\Sigma x - 2m\Sigma xy - 2c\Sigma y}{m^2+1}$$

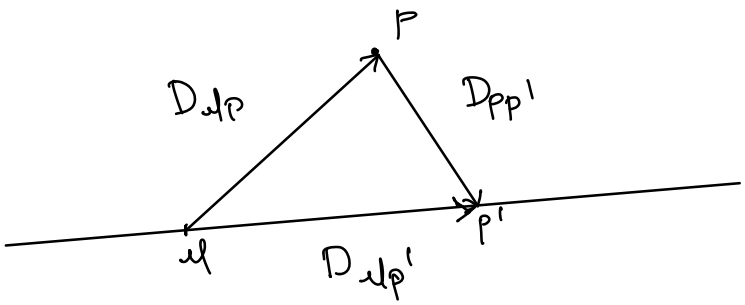
Let us first try to minimize w.r.t. c,

$$\frac{\partial e}{\partial c} = \frac{2cn + 2m\Sigma x - 2\Sigma y}{m^2+1}$$

$$\Rightarrow \Sigma y = m\Sigma x + cn$$

$$\Rightarrow \Sigma \frac{y}{n} = m \cdot \frac{\Sigma x}{n} + c$$

$\therefore$ Line passes through mean.



$$D_{elp}^2 = D_{pp'}^2 + D_{elp'}^2$$

$$\sum D_{elp}^2 = \sum D_{pp'}^2 + \sum D_{elp'}^2$$

$$Var = \sum D_{pp'}^2 + \sum D_{elp'}^2$$

Hence if we fix variance if $\sum D_{pp'}^2$ is minimized, $\sum D_{elp'}^2$ is obviously maximized to give that the line vector.

$\left(\frac{m}{\sqrt{m^2+1}}, \frac{1}{\sqrt{m^2+1}}\right)$ is the eigen vector of maximum eigen value.

$\therefore$ Proved.

Final algorithm to find the line :-

① Do the Principle component Analysis using eigen vector and values and get the vector V.

② Draw a line with directional vector as V and initial point as el.

③ That would be your best fit line.

$$C = \Sigma \begin{bmatrix} x \\ y \end{bmatrix} [x \; y]$$

$$D, V = eig(C)$$

$$\{V_{ultmax}\} \rightarrow \frac{V_{ultmax}(x)}{V_{ultmax}(y)} = \frac{1}{m}$$

$$y = mx + c$$

$$c = \Sigma \frac{y}{n} - m \Sigma \frac{x}{n}$$

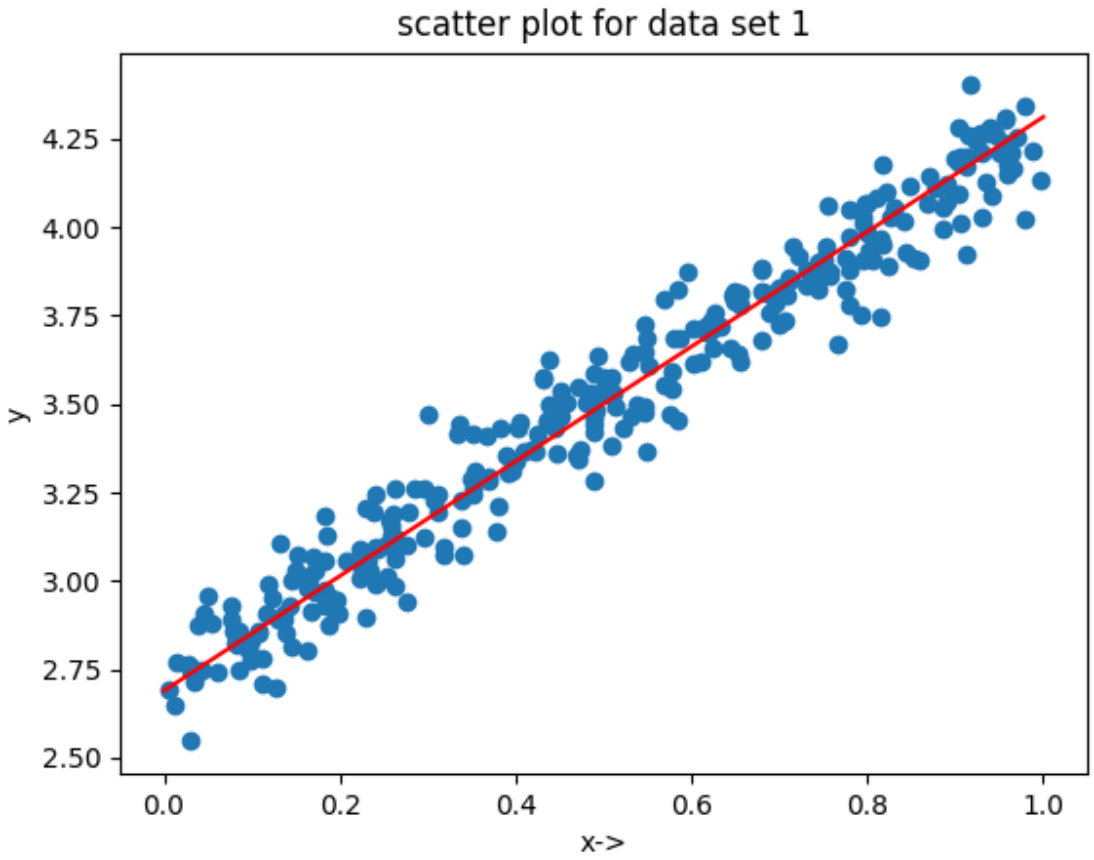
---

b.

CODE ⇒ q3_b.py GRAPH ⇒ Scatter_3b.png

The procedure given in the previous question has been applied in the given code.

Plotted the scatter plot and also the line which shows the linear relation between X and Y variables that we got along the eigen vector with maximum eigen value.



SCATTER PLOT FOR DATASET 1

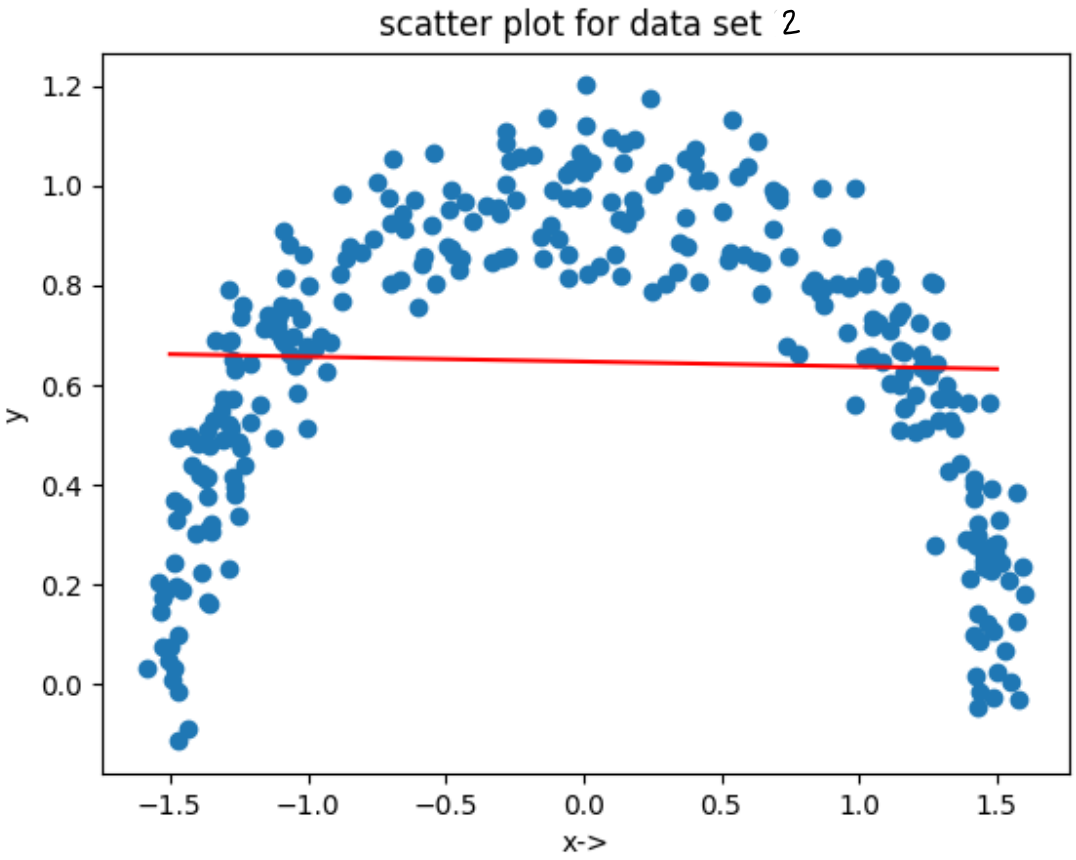
c.

CODE ⇒ q3_c.py
q3_C.ipynb

IMAGE ⇒ scatter_3C.png

OBSERVATIONS:

- Data in set 2 does not belong to variables which are related by a linear relation
- This approximation works best when data follows nearly linear trend with slight deviation and errors
- But whenever the relation exists the PCA analysis gives the relation which is very accurate.
- Like the set 1 case the scatter plot is linear (mostly) but in set 2 case the scatter plot is non-linear, so the approximate in set 2 case is not valid.
- So the linear relation we get from the PCA analysis doesn't imply that there is a relation. But if the relation exists then we can get it by this PCA analysis.


scatter plot for data set 2
1.2
1.0
0.8
0.6
0.4
0.2
0.0
y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
x->
SCATTER PLOT FOR DATASET 2