DATE

3 a) Let $\vec{\mu}$ be mean of given sample data and C be the co-variance matrix.

Let $\vec{v}$ be a unit vector along the direction in which spread is maximum, then a line passing through mean & parallel to $\vec{v}$ will give our linear approximation relationship between X & Y.

Let $\vec{v_1}$ & $\vec{v_2}$ be eigenvectors of C with eigenvalues λ_1 & λ_2.

Let $\lambda_1 > \lambda_2$ w.l.o.g

$\therefore$ Spread will be max. in direction in $\vec{v_1}$.

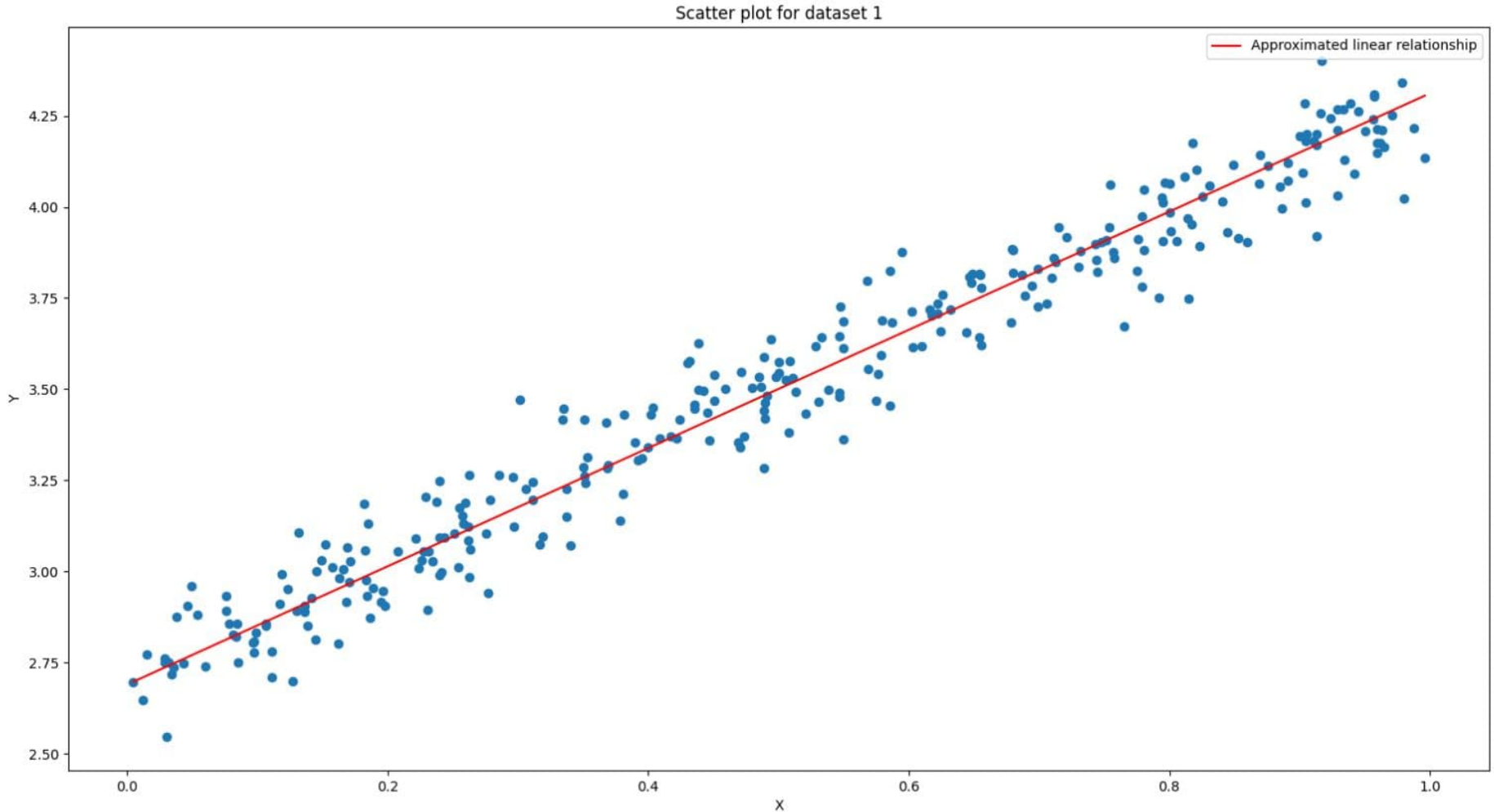
DATE

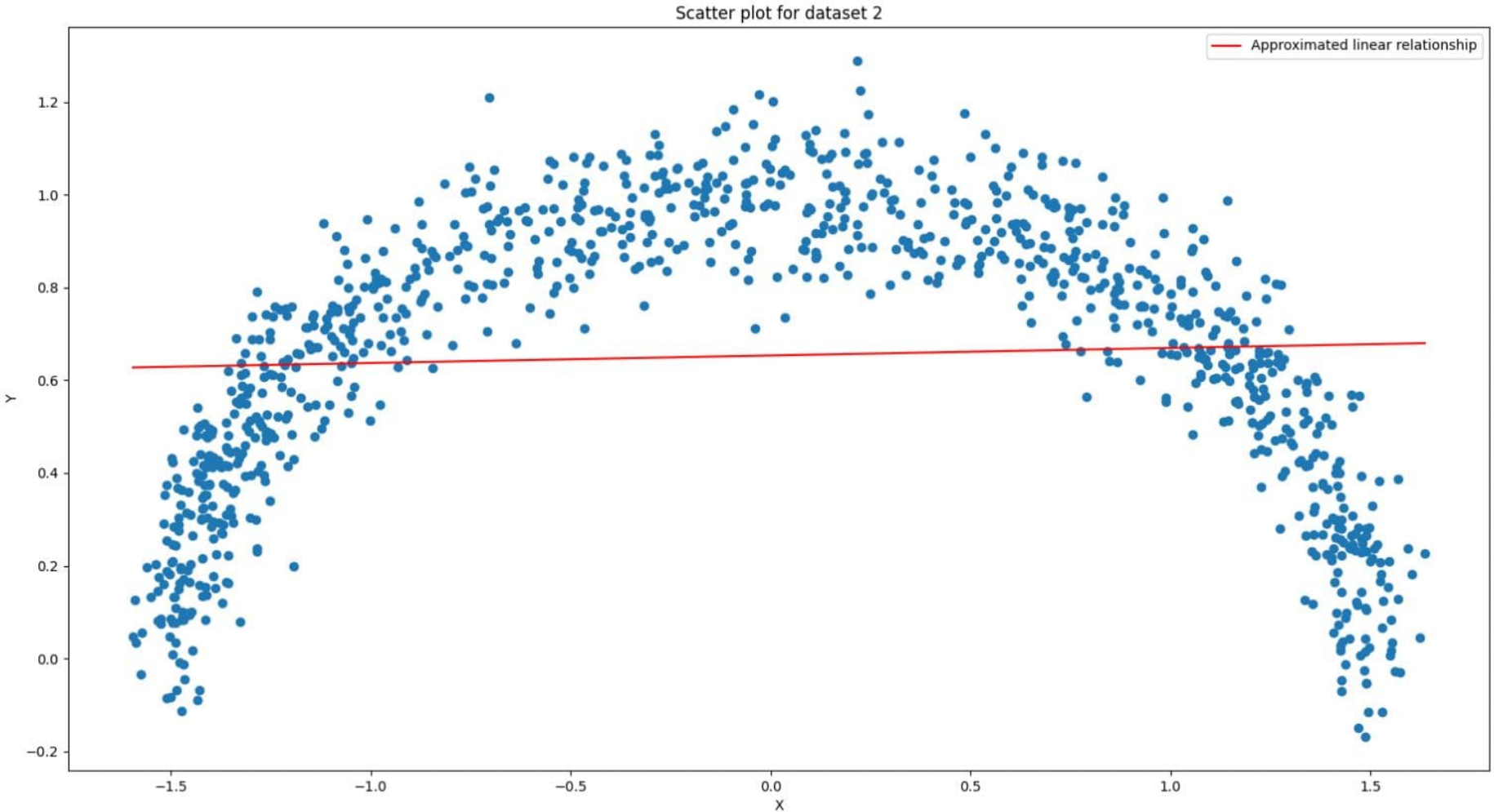
$$\therefore \vec{V} = \frac{\vec{V_1}}{|\vec{V_1}|}$$

$\therefore$ Line l given by

$$\vec{l} = \vec{a} + k\vec{V}$$

is our best approximation.


Scatter plot for dataset 1
Approximated linear relationship
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
Y
0.0
0.2
0.4
0.6
0.8
1.0
X


Scatter plot for dataset 2
Approximated linear relationship
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X

3 For first set of points,

one can infer from the scatter plot that those set of points can be nicely placed around a line as deviations are not that large from our calculated linear relationship, it is a good approximation of first set.

DATE

While for second sets of data, we can clearly infer that data points are following a curvy or non-linear pattern which is ~~very~~ not easy to put in a linear relationship hence diversions are larges and approximation is not that good compared to set 1.