PCA (Principal component Analysis).

* Data preprocession, Dimensionality Reduction
  Alternative – LDA, TSNE.

$X_1\ X_2\ X_3\ X_4 \ldots\ X_{100}$ Y

- Time complexity
- Difficult to generalise a data/relation.
- All the time all features aren't going to contribute.
- Hard to find shape of data.

Soln – Dimensionality Reduction.

DR:- Trying to convert dataset in another axis such that it retains original meaning of data.

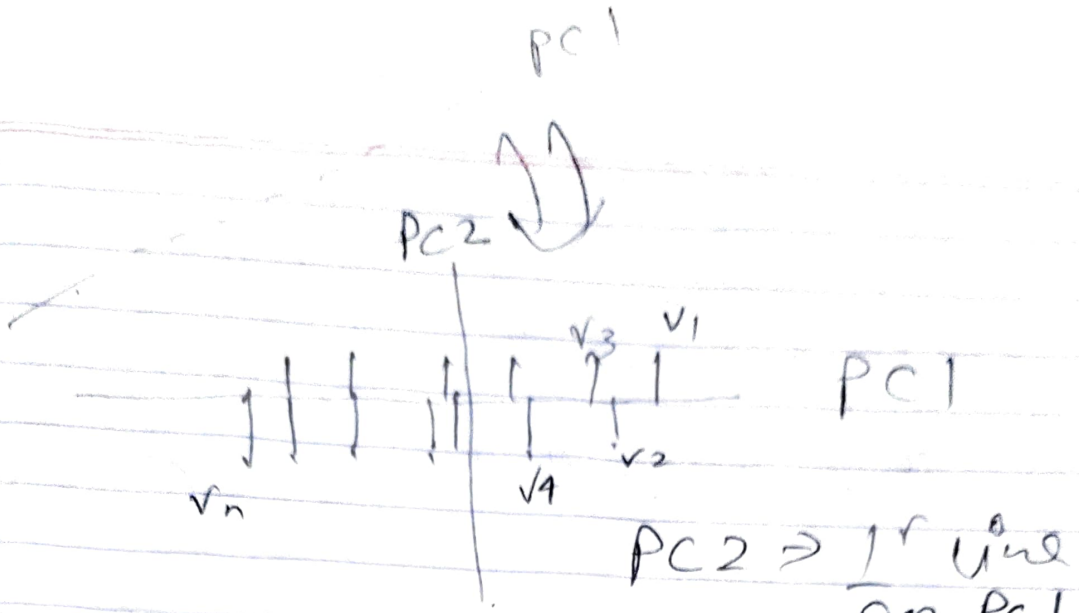
Step-1 take dataset x_1 x_2 → C
$\mu = 0$, $\sigma = 1$.

$Z = \frac{x - \mu}{SD}$ (Standard normal dist.)

Step-2 Draw single straight line explaining data.
(Implement Linear regression such that residual minimum).
(Principal component 1)

$x_2 = mx_1 + c$
or $x_1 = mx_2 + c$ Line ??

Step-3



PC2 ⇒ 1st line on PC1

New Column

V_1
V_2 ⇒ Combination of x_1 & x_2
.
.
V_n

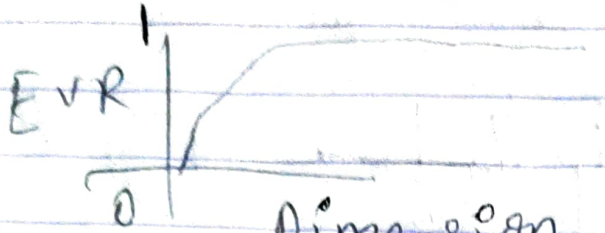
PC1 (High Priority).

How many PCR lines ⇒ EVR (Explained Variance Ratio)

EVR = How many features one PC1 line can explain?

EVR (PC1) = $\frac{\text{Distance of PC1 Points}}{\text{Distance of PC1} + \text{Distance of PC2}}$

Screen plot



$$Cov_{c_1 c_2} = \sum_{K=0}^{m-1} (c - \bar{c_1})(c_2 - \bar{c_2})$$

$m - 1$

$$\frac{A^T A}{m-1} = Cov$$

$$T(v) = \lambda v$$

$v \rightarrow$ Eigen vector

$\lambda \rightarrow$ Eigen value.

$\lambda \uparrow \Rightarrow$ PC1

np.linalg.eig(cov).

Rotation when multiplied by λ.