PCA

PCA extracts features and reduces the number of features

In some problem set, if there are 1000 of features & we try to train ML model, it will overfit

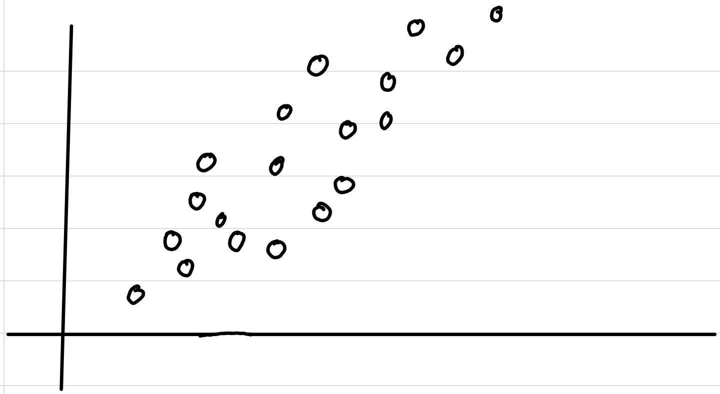
When machine is trained to take a decision, it considers features. They require space and computation for training. If we can reduce the number of features we can reduce complexity of Model

If we remove few features, we can make the model simpler

But removing features is loss of data

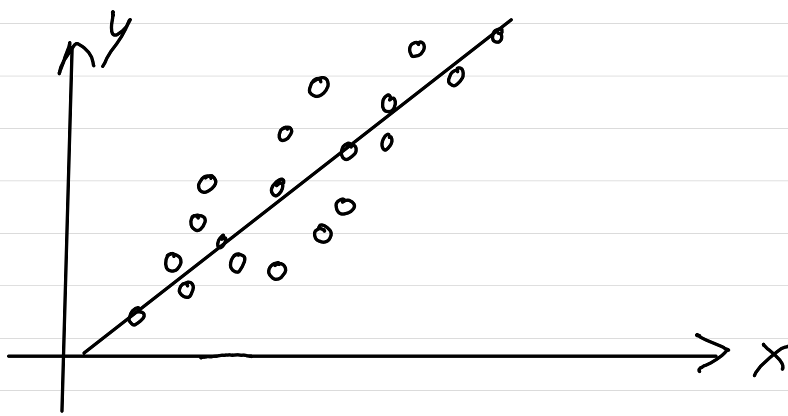
If we remove the features that are less important, we will get lower loss.

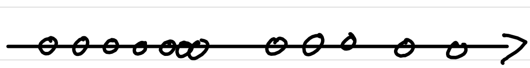
PCA automatically finds most important features to keep.



Data is of above form (2D data)

If we want to convert it into 1D data with lowest loss, we can draw a line.



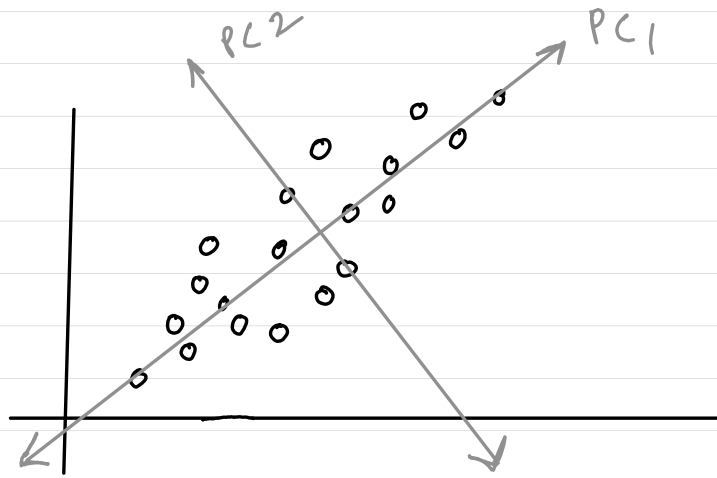


1D data representation

In order to minimize the loss we want a line that covers the maximum variation

The line that covers the maximum data points is the principle component 1

The second principle component is $\perp$ to the first.

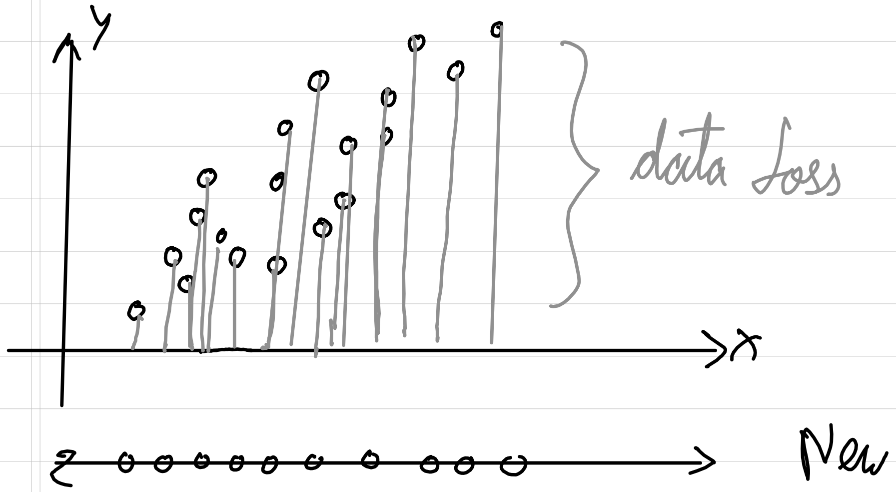


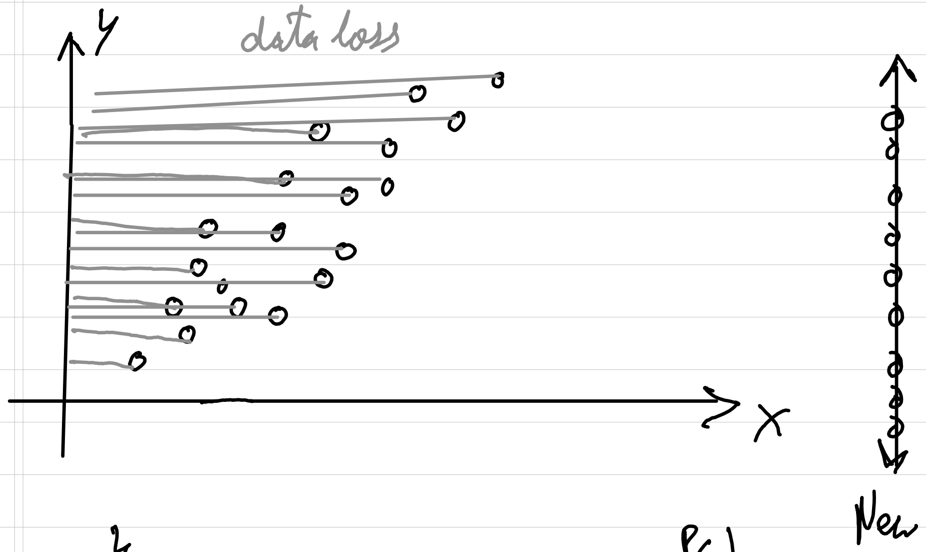
PCA is also known as K-L transform and Hotelling transform

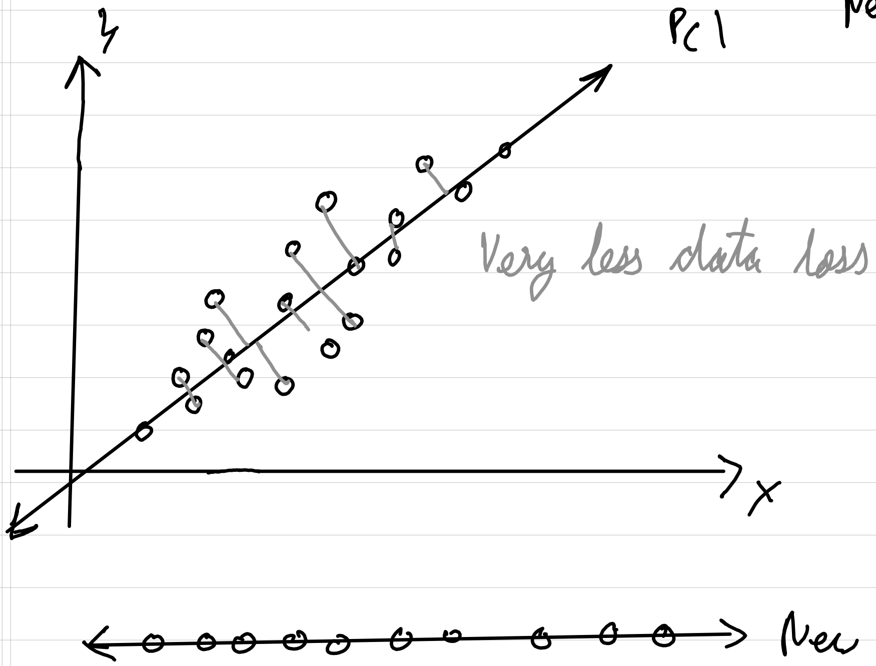
It is used in data compression

World without PCA

Why cant we drop x or y axis?







Principle components are $\perp$ axes with least loss

Steps in PCA

① standardize dataset
$$\frac{x - \bar{x}}{\sigma}$$
$$\sigma = \sqrt{\frac{(x-\bar{x})^2}{n-1}}$$

② Calculate covariance matrix
$$Cov(x, y) = \frac{\Sigma (x-\bar{x}) * (y-\bar{y})}{N}$$

③ Calculate eigenvalues & eigenvectors
$$[Cov] - \lambda I = 0$$ - solve

④ Pick k eigenvectors

⑤ Reconstruct original matrix

PCA

Principle Component Analysis

Used to identify most important points or features

(01) Data

Feature	Exp 1	Exp 2	Exp 3	Exp 4
X	4	8	13	7
Y	11	4	5	14

Reduce dimention from 2 to 1

→ step 1: No of features = 2
No of samples = 4

step 2: Calcutate Mean

$$\bar{x} = \frac{4+8+13+7}{4} = 8$$

$$\bar{y} = \frac{11+4+5+14}{4} = 8.5$$

step 3: Covariance Matrix for data

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

$$\begin{vmatrix} (x, x) & (x, y) \\ (y, x) & (y, y) \end{vmatrix}$$

Note

$$\Sigma (x - \bar{x})^2$$

$$\Sigma\, x^2 - 2x\bar{x} + \bar{x}^2$$

$$= \Sigma x^2 - 2\bar{x}\Sigma x + \bar{x}^2 N$$

$$Cov(x, x) = \frac{1}{4-1}\left((4-8)^2 + \cdots\right) = 14$$

$$Cov(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{4-1}(4-8)\times(11-8.5) + \cdots$$

$$= -11$$

$$Cov(x, y) = Cov(y, x) = -11$$

$$Cov(y, y) = \frac{1}{N-1}(\Sigma \cdots\cdots) = 23$$

Matrix $\begin{vmatrix} (x,x) & (y,y) \\ (y,x) & (y,y) \end{vmatrix} = \begin{vmatrix} 14 & -11 \\ -11 & 23 \end{vmatrix}$

step 4: Eigenvalue, eigenvector

$\lambda \Rightarrow$ eigenvalue
$V \Rightarrow$ eigenvector

$$SV = \lambda V$$

$$(S - \lambda I)V = 0$$

$$\det(S - \lambda I) = 0 \qquad \text{or } V = 0$$

$$\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\det \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = (14-\lambda)(23-\lambda) - 121 = 0$$

$$(14-\lambda)(23-\lambda) = 121$$

$$322 - 37\lambda + \lambda^2 = 121$$

$$\lambda^2 - 37\lambda + 201 = 0$$

Solving for λ

$$\lambda_1 = 30.384$$
$$\lambda_2 = 6.615$$

Consider highest value of λ

First principle component $\lambda_1 = 30.381$

eigenvector for the principle component

$$(S - \lambda I)V = 0$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda)V_1 + (-11)V_2 = 0 \quad - (1)$$
$$(-11)V_1 + (23-\lambda)V_2 = 0$$

From (1)
$\lambda = 30.38$

$$\frac{V_1}{V_2} = \frac{11}{14 - 30.38} = \frac{11}{-16.3848}$$

$$V_1 = 11$$
$$V_2 = -16.3848$$

Normalized eigen vectors

$$V = \begin{bmatrix} \frac{11}{\sqrt{11^2 + (16.38)^2}} \\ \frac{-16.38}{\sqrt{11^2 + (16.38)^2}} \end{bmatrix}$$

$$V = \begin{bmatrix} 0.557 \\ -0.830 \end{bmatrix}$$

step 5: Derive 2 to 1 using $1^{st}\lambda$

$$[V_1\ V_2] \begin{bmatrix} x_1 - \bar{x} \\ y_1 - \bar{y} \end{bmatrix}$$

$$[0.5574 \quad -0.8303] \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$P_{11} = -4.3052$$

$$(V_1\ V_2) \begin{bmatrix} x_2 - \bar{x} \\ y_2 - \bar{y} \end{bmatrix}$$

$$P_{12} \quad [0.5374 \quad -0.8303] \begin{bmatrix} 0 \\ -3.5 \end{bmatrix} = +3.7386$$

$$P_{13} = 5.69$$
$$P_{14} = -5.1238$$

Final Answer

$$[-4.3052 \quad 3.7386 \quad 5.6928 \quad -5.1238]$$