Principal Component Analysis (PCA) [Dimensionality Reduction]

⇓

① Curse of Dimensionality → Dimensions

Dataset : 500 features

Price of the House

- House size
- No. of bedrooms
- No. of bathrooms

3 features → M1 → Acc 1

6 features → M2 → Acc 2 ↑↑

15 features → M3 → Acc 3 ↑↑↑

50 features → M4 → Acc4 ↓↓ ← Model is overfeeded

Acc 1 > Acc 2 ↑↑ > Acc 3 ↑↑↑ < Acc4 ↓↓

100 features → M5 → Acc 5 ↓↓↓

500 features → M6 → Acc 6 ↓↓↓↓

< Acc 5 ↓↓↓ <

② Model performance Degrade

Domain

Person ↓↓↓

⇑

Confused   Accurate

f1   2BHK

Loca ← 450K - 500K

3BHK ← 500K - 600K

Beach ← ↑↑↑

Near to a celebrity house ← ↑↑↑

Grocery shop ← ↑↑

School ←

Original

f1 f2 f3 O/p

Lesser ⇓ Feature Extraction

D1 D2 O/p

Two different ways to remove curse of Dimensionality

① Feature Selection ⇑

② Dimensionality Reduction (PCA) ↓↓↓↓ ⇓

Imp features

Feature Extraction

Feature Selection Vs Feature Extraction

↳ Dimensionality Reduction

① Why Dimensionality Reduction?

3d | 2d

100d ⇓ 3d or 2d

- Prevent → Curse of Dimensionality
- Improve the performance of the model
- Visualize the data → understand the data

Feature Selection

I/P	O/P
X	→ y
–	–
–	–
–	–
–	–

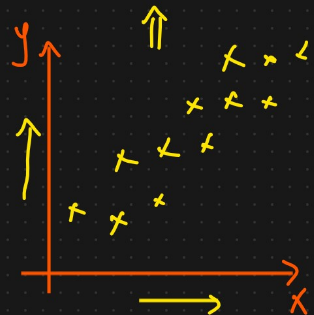
+ve ⇒

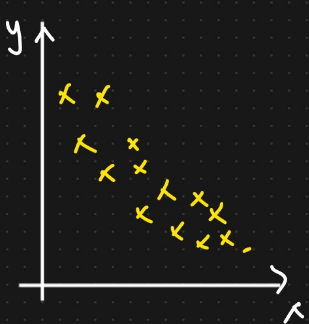
X↑	y↑
X↓	y↓

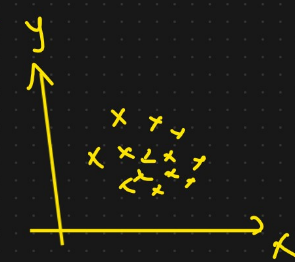
X↓	y↑
X↑	y↓

⇐ −ve

No relationship between X & Y







$$\underline{Cov(x,y)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{N-1} \begin{array}{l} = +ve \\ \approx 0 \\ = -ve \end{array}$$

≈ 0 {No Relationship}

$\boxed{-ve \text{ correlated}}$

$$\text{Pearson Correlation} = \frac{Cov(x,y)}{\sigma_x * \sigma_y} = \boxed{-1 \text{ to } 1}$$

The more toward the value of +1 the

more +ve correlated X & Y is

Dataset Housing

Independent feature

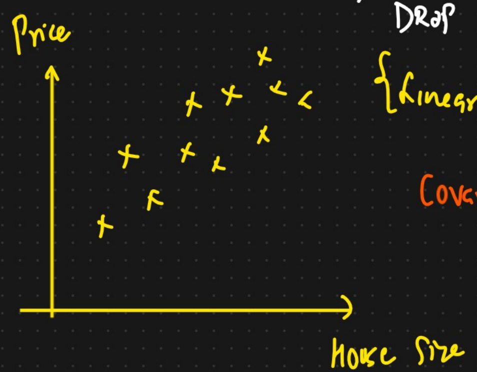
✓ House Size | Fountain Size | Price

X

o/p feature

Drop

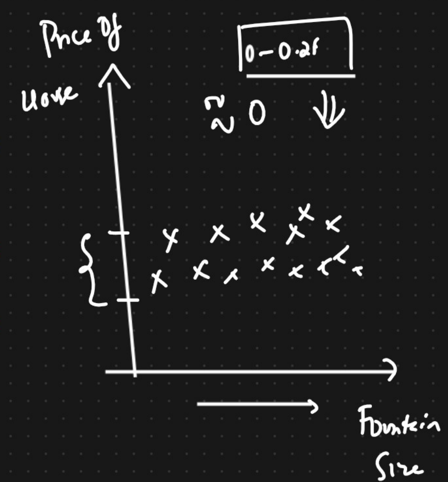
{Feature Selection}



{Linear Relationship}

Covariance = +ve

-ve



Domain Expertise

Feature Extraction

+ve

-ve

2 feature → 1 feature

Dimensionality Reduction

{ Room Size | No. of Rooms | Price | Price

⇓ ⇓ Transformation To extract New feature

→ House Size | Price

PCA Geometric Intuition

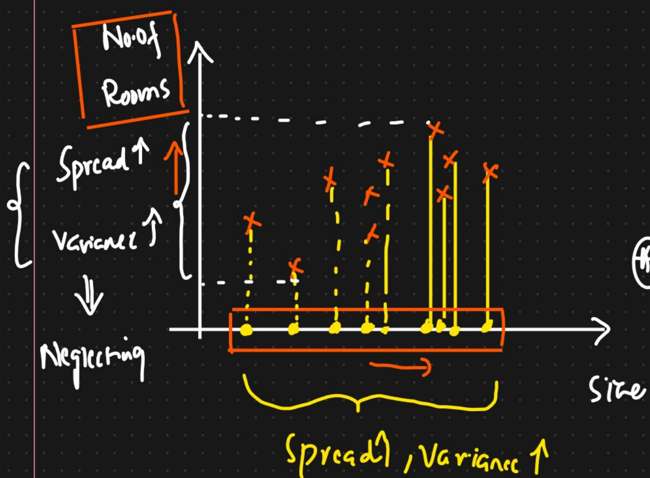
{Dimensionality Reduction}

Housing Dataset

Size House	No. of Room	Price ($\nearrow$ O/P)

PCA

$\boxed{\text{2 dimension} \longrightarrow \text{1 dimension}}$

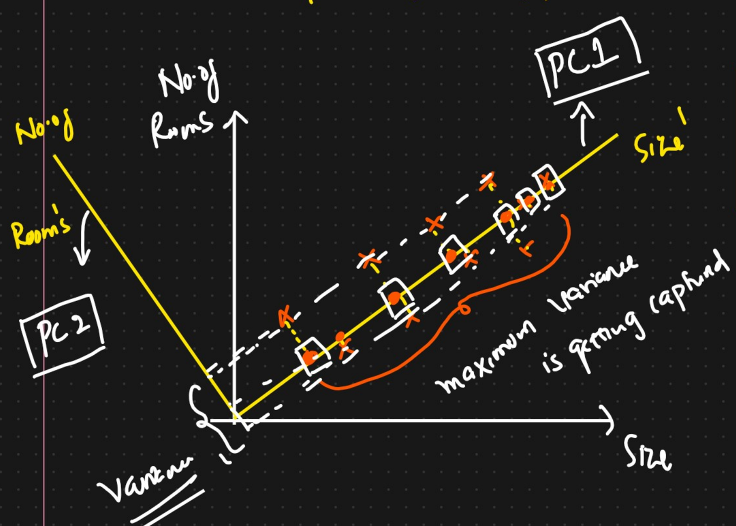


2D → 1D

(+) Loss of information (No. of Rooms)

⇓

Feature Extraction



$\boxed{\text{Eigen decomposition on Matrix}}$

⇑

Transformation

2 Dimension

PC1, PC2

Var(PC1) > Var(PC2)

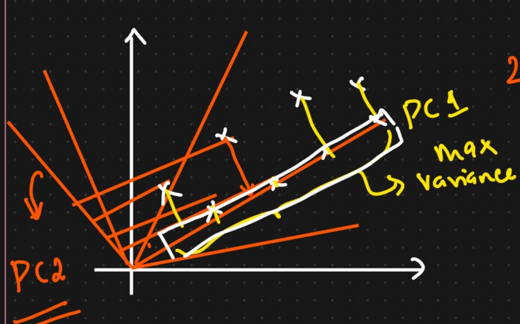
3 Dimension

PC1, PC2, PC3

Var(PC1) > Var(PC2) > Var(PC3)

2D → 1 Dimension

much Information is not lost



2D → 1D

2 Best Principal Component

To get the best Principal Component which captures maximum variance

$3D \rightarrow \boxed{1D}$

$3D \rightarrow 2D$

$\Downarrow$

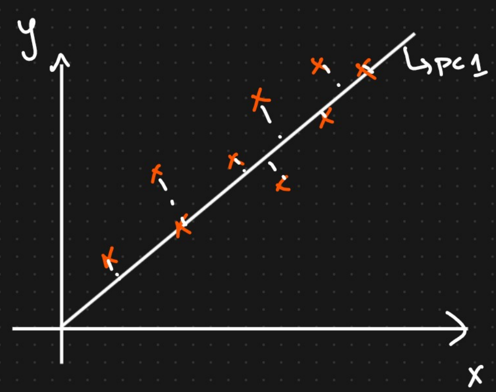
$\boxed{PC1}$, $\boxed{PC2}$, $PC3$

PC1 → 1 Dimension

PC1, PC2 → 2 Dimension.

$Var(PC1) > Var(PC2) > var(PC3)$

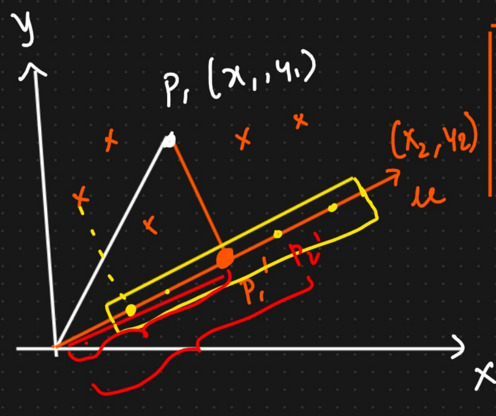
Maths Intuition behind PCA Algorithm



$2D \rightarrow 1D$

① Projection ✓

② Cost function → Variance ✓



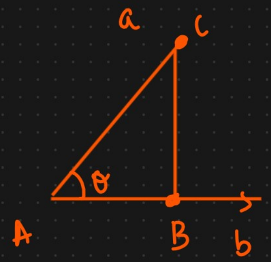
$$\boxed{Proj_{P_1} u = \frac{P_1 \cdot u}{\|u\|}}$$

$\|u\| = 1 \rightarrow$ unit vector

$$\boxed{Proj_{P_1} u = P_1 \cdot u} \Rightarrow \text{Scalar value}$$

$\Downarrow$

P_1'



$$\boxed{P_0', P_1', P_2', P_3', P_4', \ldots P_n'}$$

$\Downarrow$

Scalar values

$\Downarrow$

Variance

$$\boxed{P_0', P_1', P_2', P_3', P_4', \dots P_n'}$$

$\Downarrow$

$x_0', x_1', x_2', x_3', x_4' - - x_n'$

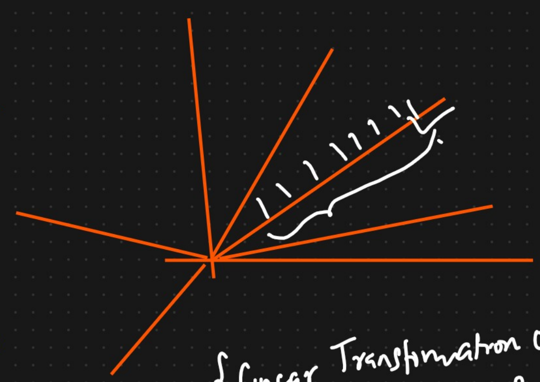
$$\text{Max Variance} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

{Goal: Find the best unit vector which captures maximum variance}.

$\Downarrow$

Cost function

Eigen vectors And Eigen values.



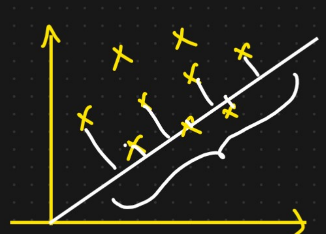
1. Covariance Matrix between $\boxed{\text{features}}$
2. Eigen vectors and Eigen values will found out from this covariance matrix

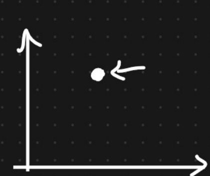
{Linear Transformation of matrix}

$$\boxed{Av = \lambda v}$$

3. Eigen vector ⟶ Eigen value → magnitude of the Eigen vector ⟶ Capture the maximum Variance

Eigen vectors And Eigen values [Linear Transformation]



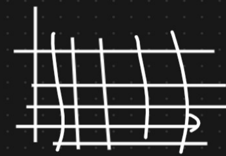


[Eigen decomposition of covariance Matrix]

$\Downarrow$

Eigen vector & Eigen values

$$\begin{bmatrix} & & \end{bmatrix} * [v] = \underset{\substack{\downarrow \\ \text{Eigen} \\ \text{Value}}}{\lambda} * v$$



$$\boxed{A * v = \lambda * v}$$

$\Uparrow \quad \Downarrow$

Eigen vector → Maximum magnitude

⇓

Principal Component

⇓

Max Var

Eigen vector → Max Magnitude

⇓

Max Eigen vector

⇓

Best Principal Component → PC1

Steps to calculate Eigen value and vectors

① Covariance of features

$\boxed{X, Y}$ $\quad Z$

↓

X'

$$Cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

2x2

	X	Y
X	Var(X)	Cov(X,Y)
Y	Cov(Y,X)	Var(Y)

A =

$Cov(X,X) = Var(X)$

$Cov(Y,Y) = Var(Y)$

	X	Y	Z
X			

$\boxed{\lambda_1, \lambda_2, \lambda_3}$

$\boxed{f_1 \text{ \& } f_2}$

$$\boxed{A \cdot v = \lambda \cdot v}$$

$\boxed{\lambda_1}$ $\boxed{\lambda_2}$

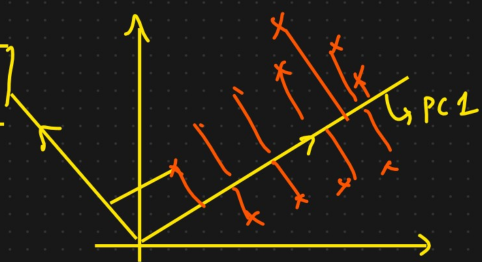
$\Downarrow$ $\Downarrow$

PC1 PC2

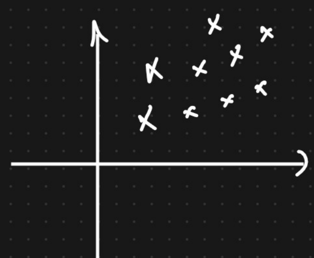
$\boxed{\lambda_1, \lambda_2}$ → Eigen values

$\Downarrow$

$\boxed{\text{PC1, PC2}}$

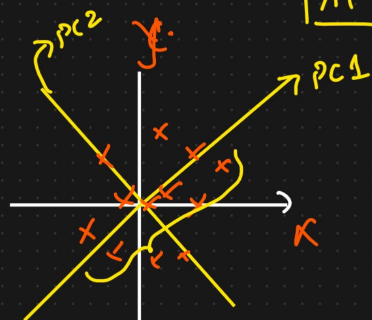


4



2D → 1D

$\boxed{\lambda_1}$ ⇒ magnitude of the Eigen vector



1 Standardize the data

2 Covariance Matrix of X & Y

$$A = \begin{array}{c|cc} & X & Y \\ \hline X & Var(X) & Cov(X,Y) \\ Y & Cov(Y,X) & Var(Y) \end{array} \quad 2\times2$$

3 Find out Eigen vectors And value

$$Av = \lambda v$$

$\boxed{\lambda_1, \lambda_2}$ ⇒ Eigen values.

↓ ↓

PC1 PC2

$3D \to 2D$

$\lambda_1, \lambda_2, \lambda_3$

$\downarrow \quad \downarrow \quad \downarrow$

PC1 PC2 PC3

$\boxed{1D}$ $\boxed{1D}$ $\Rightarrow 2D$

$3D \to 1D$

$\lambda_1, \lambda_2, \lambda_3$

$\boxed{1D}$

$2D \to 1D$

λ_1, λ_2

$\downarrow \quad \downarrow$

PC1 PC2

$\boxed{1D}$ $\to$ $\boxed{\text{Projection}}$