Principal Component Analysis (PCA)

[Dimensionality Reduction]

→ Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in machine learning and data analysis

→ Its primary goal is to reduce the dimensionality of a dataset while preserving as much of the variability (information) as possible.

→ This can be particularly useful when dealing with high-dimensional data, as it helps in visualizing and understanding the underlying structure of the data, as well as in speeding up computation and reducing noise.

500 features → ML model → ACC ↓↓

↓

Information capture

↓ ↓ ↓

f1 f2 f3 → 3D

⇓ PCA

2D

2D → 1D

PCA → Feature Extract → Dimensionality Reduction

[Feature Selection]

⇓ 3D

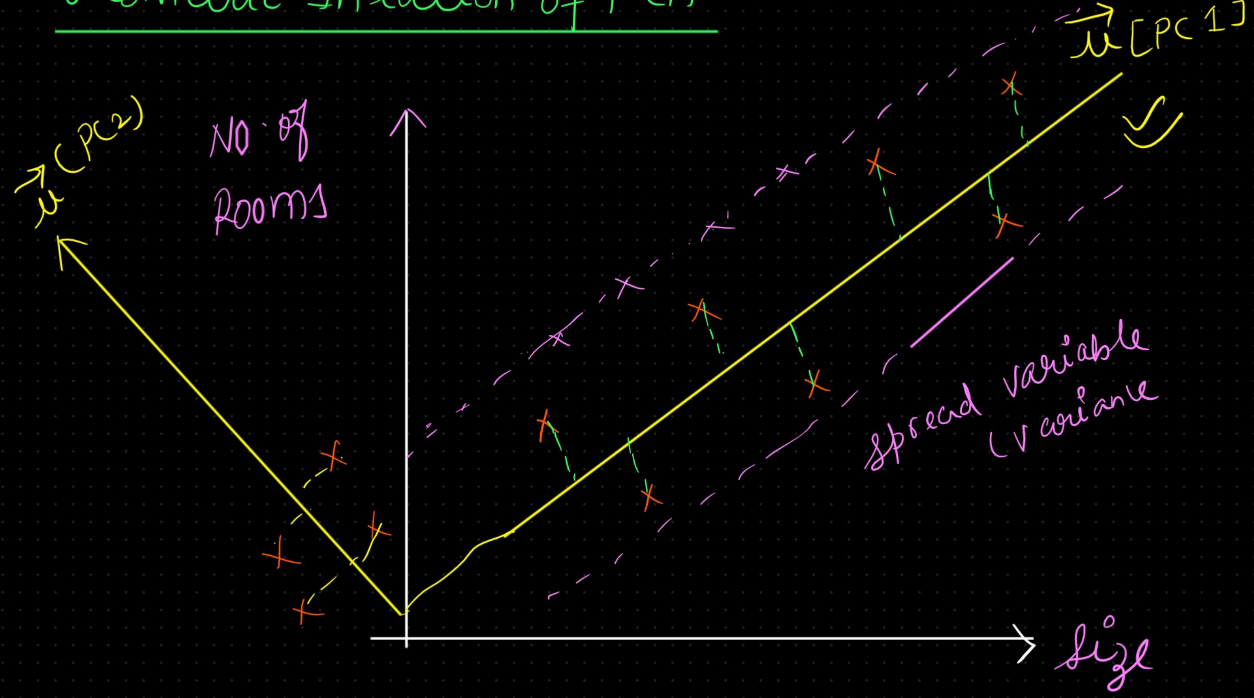
[Max. Variance/ Spread]

↓

2D

∴ Model → ACC↑↑

Geometric Intution of PCA



2 features → PC1, PC2

↓

Max variance

Vector → magnitude
→ direction

3 features → $\underbrace{\text{PC1, PC2}}$, PC3

max variance/

Spread

[Information of the Data]

Mathematics Intution

1. Covariance Matrix

x, y

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

$$A = \begin{array}{c|cc} & x & y \\ \hline x & \operatorname{Var}(x) & \operatorname{Cov}(x,y) \\ y & \operatorname{Cov}(y,x) & \operatorname{Var}(y) \end{array}$$

2. Eigenvalues and Eigen vectors

magnitude → Eigenvalues [variance value]

dir → Eigen vectors [Max variance]

Eigenvalues and Eigenvectors: PCA then calculates the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions (principal components) of maximum variance in the data, while eigenvalues represent the magnitude of variance along those directions.

$$\boxed{A V = \lambda V}$$

$A \rightarrow$ square matrix (cov-matrix)

$V \rightarrow$ eigen vector

$\lambda =$ eigen value

(3) Dimensionality Reduction

We select the top eigenvector (principal component) that captures the most variance in the data. This eigenvector represents the direction along which the data varies the most. By projecting our data onto this eigenvector, we reduce the dimensionality from 2D to 1D while retaining the maximum variance possible.