# Principal Component Analysis (PCA)

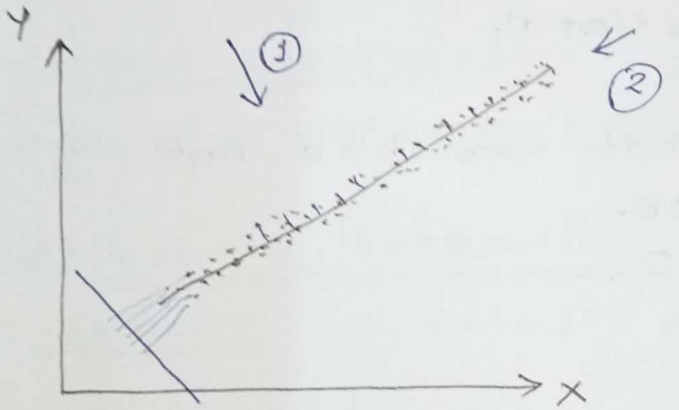
## PCA :-

↳ Unsupervised learning algo used for Dimensionality Reduction in ML.

↳ Solve overfitting (having many useless features).

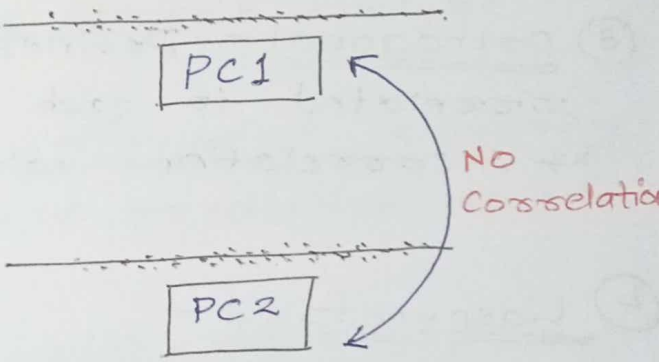
↳ Principal Components: statistical process convert observations of correlated features into set of linearly uncorrelated features with help of orthogonal transformation. New features c/a PC.

↳ PCA is used to find PC (Principal component):




consider Imaginary Line

Reduce 2D to 1D.




- Orthogonal: No correlation bet$^n$ different Principal components.

↳ So, num. of PC <= No. of features.

↳ When form large num. of PC's give most importance to PC 1 (First PC).

## ⊙ Math concepts used in PCA :-

① Variance & covariance
② Eigen values & Eigen vectors.

## ⊙ Common Terms used in PCA Algo :-

① **Dimensionality** – Num. of features / variables in data.
↳ Same as num. of columns in dataset.

② **correlation** – How strongly 2 variables related.
↳ i.e. if one changes ⇒ other var also changes.
↳ Value ranges between −1 to +1.
−1: Inversely proportional.
+1: Directly proportional.

③ **Orthogonal** – Defines that variables not correlated to each other.
↳ ∴ correlation value 0.

④ **Eigenvectors** – Represent the direction of PC.

⑤ **Eigenvalues** – Repbesent variance explained by each PC.

⑥ **Covariance Matrix** – Matrix contain covariance betⁿ pair of variables.

- **Principal Components in PCA :**

↳ Transformed new feature / o/p of PCA is cla PC.

↳ Num. of | PC <= num. of features | in dataset.

↳ Pc must be linear combination of original features.

↳ components are orthogonal i.e. 0 correlation.

↳ PC1 is most important & decreases as move to PCn.

- **Uses of PCA :**

① Data compression – Reduce dimensionality of high-dim. dataset, easy to store & analyze.

② Feature Extraction – Identify most important features in dataset & build predictive model.

③ Visualization – Visualize high-dim data in 2D. OR 3D, make easy understand & interpret.

④ Data Pre-processing – can use as Pre-proc for other ML algo( clustering & classification).

## Applications of PCA:

(1) Dimensionality Reduction in computer vision, Image compression.

(2) Find Hidden Patterns in data with high-dim. Used in Finance, data mining, Psychology.

## Advantages of PCA:

(1) Dimensionality Reduction - Reduce num. of variables.

(2) Feature Selection - Select most imp features. Identify imp features, from large num. of variables.

(3) Data Visualization - By reducing num. of features, plot high-dim. data in 2D/3D, easy to interpret.

(4) Multicollinearity - Problem of 2/3 highly correlated variables. Identify structure in data & create new, uncorrelated variables used in Regr$^n$ model.

(5) Noise Reduction - Reduce noise in data. Remove PC with low variance (Assumed to have noise).

(6) Data Compression - Represent data using smaller num. of PC, which capture most of variation in data, PCA reduce storage requirements & speed up processing.

(7) Outlier Detection - Outliers different from other points. PCA identify outliers from data, that are far from other points in PC space.

## ⊙ Limitations of PCA :

① Interpretation of PC – PC are linear combination of original variables, difficult interpret with original variables. Difficult to explain results of PCA.

② Data Scaling – PCA sensitive to data scaling. Imp to scale data before applying PCA.

③ Information Loss – PCA reduce num. of variables, can lead to loss of information. Degree of Info. loss depend on num. of PC used. so, more num. of PC ⇒ More Information loss.

④ Non-Linear Relationships – PCA assum linear relationship betⁿ variables. so, if non-linear relationships ⇒ PCA not work properly.

⑤ Computational Complexity – For large datasets. Especially for large num. of variables.

⑥ Overfitting – when model fits training data too well & perform poor on new data. Can happen, if too many PC used & train on small data.