Principle Component Analysis (PCA)

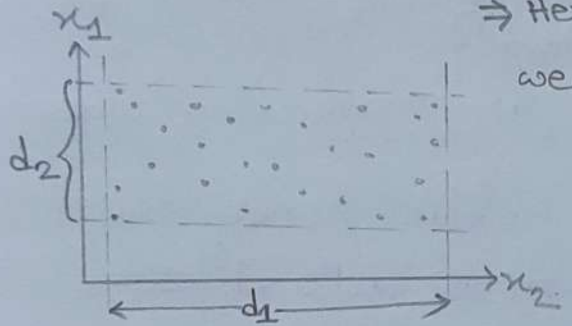
PCA is a feature extraction technique that tries to convert the higher dimensional data to a lower dimension while keeping the actual behaviour.

Benifits of PCA

① Reduce the dimension → Faster execution of model.

② visualization.

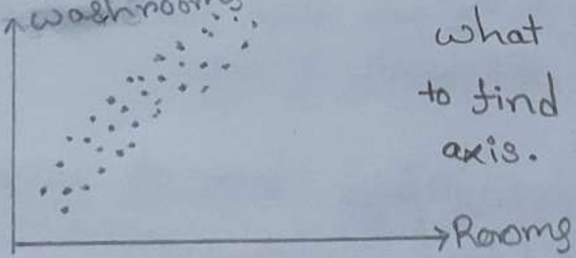
lets we have 2 col's and we have to choose one. Then we will take which variance (spread) is more



⇒ Here x_2 has more variance so we will take x_2.

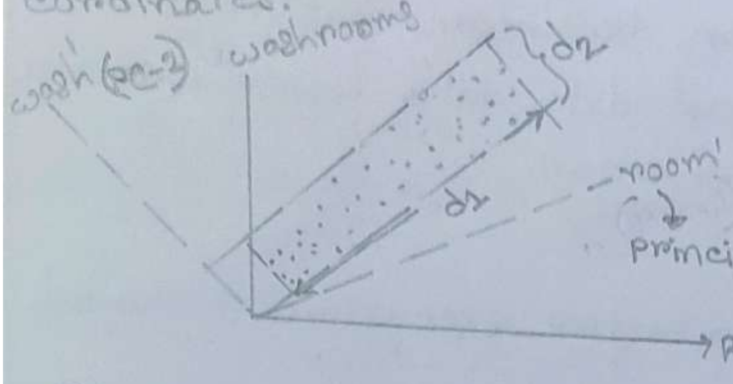
Problem is what if x_1 and x_2 has almost same variance. At that can feature selection fails. And we use feature extraction.

PCA takes the dataset and recreate features from it. Then give top most important features.



what PCA will do is: It will try to find a new set of cordinates axis.

PCA will rotate the axixs and try to find some new cordinates.



principle component-1

as in new axis $d_1 > d_2$

So we will take room' → PC1

So now we will take Room and transform this into PC-1 (room')

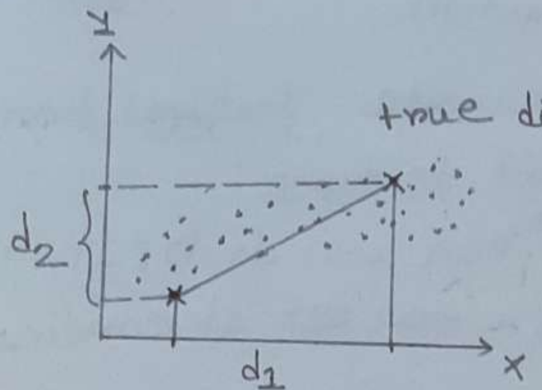
Note: no. of principle components $\leq n$ → no of original feature.
and each compnent is perpendicular with others

Why variance (spread) is important?

· variance is not spread it's propotional to spread.

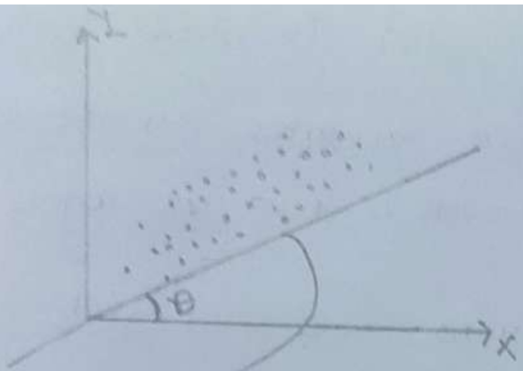
In PCA we don't use mean Absolute error cause absolute functions are not differentiable at zero

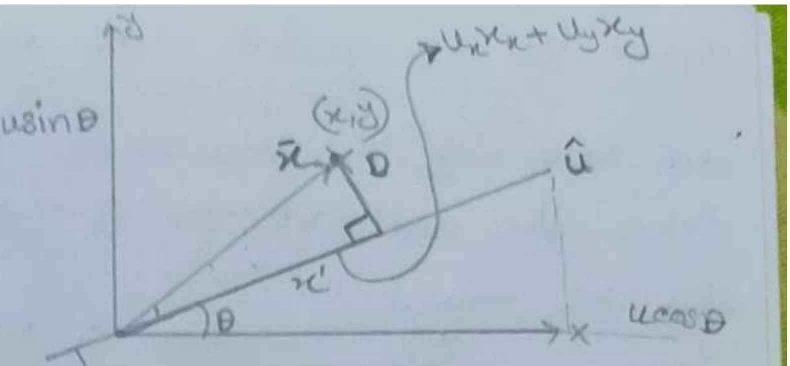
$$MAD = \frac{\sum |x_i - \bar{x}|}{N} \qquad variance = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$



true distance between these 2 points was d1 but if we project this on "y" axis then the actual distance is lost.

so to prevent the sparsity and lost of actual distance we maintain a large variance.

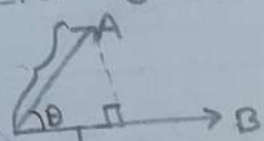




It's a vector. It can be any length (we don't care) so we take it as unit vector → let's say unit vector $\hat{u}$

Projection of vector $\bar{A}$ on vector $\bar{B}$ is

$$\text{Proj}_{\bar{B}}\bar{A} = \frac{\bar{A}\cdot\bar{B}}{|B|}$$



$\boxed{A\cos\theta}$ → projection

$\Rightarrow |A|\cos\theta \times \frac{|\bar{B}|}{|\bar{B}|}$

$\Rightarrow \frac{A.B}{|B|}$

Projection of vector $\bar{x}$ D on $\hat{u}$ vector

$$x' = \frac{\bar{x}\cdot\bar{u}}{|\bar{u}|} = \bar{x}\cdot\bar{u} \quad \begin{bmatrix}\text{cause} \\ |u| = 1\end{bmatrix}$$
$$= \bar{u}^T . x$$

$\hat{u}$ has 2 component u_x, u_y and $\bar{x}$ also has 2 comp. x_x, x_y

$$\therefore u^T . x = \begin{bmatrix} u_x & u_y \end{bmatrix}^T . \begin{bmatrix} x_x & x_y \end{bmatrix}$$

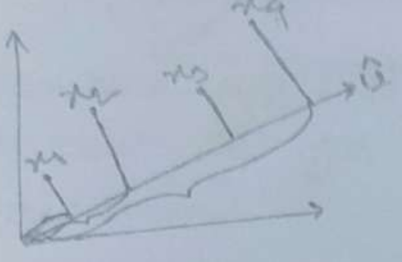
$$= \begin{bmatrix} u_x \\ u_y \end{bmatrix} . \begin{bmatrix} x_x & x_y \end{bmatrix}$$

For a single point

$$= \boxed{u_x x_x + u_y x_y}$$ → a scalar value

Same for n points we will get n scalar values.

our task was to maximize the variance. so we will take the unit vector $\hat{u}$ for which our variance will be max.



For n points we will get

$$\left[u^T x_1 \quad u^T x_2 \ldots\ldots u^T x_n \right]$$

now we have to find the variance of these points.

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

$$\boxed{Var = \frac{1}{n}\sum_{i=1}^{n}\left(u^T x_i - u^T \bar{x}\right)^2}$$

$u^T\bar{x} \longrightarrow \bar{x}$ is the mean of all points. And $u^T\bar{x}$ is the value of $\bar{x}$'s projection on $\hat{u}$

now find a unit vector u for which var will be maximum.

$$var(u) = \frac{1}{n}\sum_{i=1}^{n}\left(u^T x_i - u^T \bar{x}\right)^2$$ where $\|u\| = 1 \longrightarrow$ unit vector

let $X = [x_1 \; x_2 \; \ldots\ldots x_n]^T \in \mathbb{R}^{n\times d}$.

$\bar{x}$ = mean of the data $\bar{x} \in \mathbb{R}^d$ a column vector

$$X_{centered} = X - \bar{X}$$

$\boxed{\text{step-1}}$ $Z_i = u^T(x_i - \bar{x})$ so $var(u) = \frac{1}{n}\sum_{i=1}^{n}(Z_i)^2$

$$var(u) = \frac{1}{n}\sum_{i=1}^{n}\left\{u^T(x_i - \bar{x})\right\}^2$$

here u is a column matrix $[u \in R^{d\times 1}]$

$$u^T(x_i - \bar{x}) \longrightarrow (d\times 1)^T . (d\times 1)$$
$$= (1\times d)(d\times 1)$$
$$= (1\times 1) \longrightarrow \text{a scalar}$$

$\boxed{\text{Note}}$

$x_i \in R^{d\times 1} \longrightarrow$ a single point is a column matrix with d features.

$\bar{x} \in R^{d\times 1} \longrightarrow$ also a column matrix cause d number of mean for d no of features.

so projection direction u is also a column matrix

$\therefore u^T \in R^{1\times d} \longrightarrow$ a row vector

u is a vector and all vector are column matrix

$\boxed{\text{step-2}}$ $u^T(x_i - \bar{x}) \in \mathbb{R} \longrightarrow$ a scalar

now $\left\{u^T(x_i - \bar{x})\right\}^2 = \left[(x_i - \bar{x})^T . u\right] . \left[u^T(x_i - \bar{x})\right]$

$$= (x_i - \bar{x})^T (uu^T)(x_i - \bar{x})$$

$$var(u) = \frac{1}{n}\sum_{i=1}^{n}\left\{(x_i - \bar{x})^T . (u.u^T) . (x_i - \bar{x})\right\}$$

$\boxed{\text{step-3}}$ we have total n data points

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in R^{n\times d}$$

vector mean of X

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \left[x_i \in R^{d\times 1}\right]$$

let

$$Z = \begin{bmatrix} Z_1^T \\ Z_2^T \\ \vdots \\ Z_n^T \end{bmatrix} = \begin{bmatrix} (x_1-\bar{x})^T \\ (x_2-\bar{x})^T \\ \vdots \\ (x_n-\bar{x})^T \end{bmatrix} \in R^{n\times d}$$

each row of Z is now centered. (mean $=0$)

$\boxed{\text{Goal}}$

$$Var(u) = \frac{1}{n}\sum_{i=1}^{n}\left(u^T Z_i\right)^2$$

$$\# \; (U^T Z_i)^2 = (Z_i^T . U)(U^T . Z_i) = \underbrace{Z_i^T u u^T Z_i}_{\text{Scalar}} = \text{trace}(U^T Z_i Z^T U)$$

[Note: scalar = trace(scalar)]

now,

$$Var(u) = \frac{1}{n}\sum_{i=1}^{n} tr\left(U^T Z_i Z_i^T U\right)$$

$$= tr\left(U^T\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T\right)U\right)$$

now define the covariance matrix:

let

$$S = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T = \frac{1}{n} Z^T Z \quad \left[(1\times d).(d\times 1) = d\times d\right]$$

now $Var(u) = tr(u^T S u) = u^T S u$

$u \in \mathbb{R}^{d \times 1} \rightarrow$ a col matrix

$S = Z^T . Z \Rightarrow S^{d \times d}$ a square matrix (symmetric)

$$u^T S u = \underbrace{(1 \times d)\ (d \times d)}\ (d \times 1)$$
$$= (1 \times d) \cdot (d \times 1)$$
$$= (1 \times 1) \quad \therefore u^T S u \text{ is a scalar}$$

$tr(scalar) = Scalar$

$\therefore tr(u^T S u) = u^T S u$

For (2x2) data covariance matrix $= \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & cov(x_2, x_2) \end{bmatrix}$

$\boxed{cov(a, a) = var(a)}$

$\boxed{cov(a, b) = cov(b, a)}$

$$= \begin{bmatrix} var(x_1) & cov(x_1, x_2) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$$

2x2 square symetric matrix

For (nxn) things will be same

$var(u) = u^T S u$ to find the unit vector u so that $var(u)$ will be max we have to find the eigen vector and eigen value of S.

S is a square matrix shape $(d \times d)$

From the characteristic equation

$$Sv = \lambda v$$

$$\begin{bmatrix} \lambda \rightarrow \text{eigen value} \\ v \rightarrow \text{" vector} \end{bmatrix}$$

$$\Rightarrow Sv - \lambda v = 0$$

$$\Rightarrow (S - \lambda I) v = 0$$

It's a homogeneous system. To get a nontrivial solution so $u \neq 0$ is

$\boxed{\det(S - \lambda I) = 0}$ after solving this we will get some eigen values (λ)

By taking the max value of λ and we will get a eigen vector.

From Rayleigh quotient: The max value of $u^T S u$ occurs when u is the eigen vector of S corresponding to its largest eigen value λ

Final Steps:

① Find the centered data. Z where $Z_i = X_i - \bar{X}$

② Compute Covariance matrix $\left[S = \frac{1}{n} Z \cdot Z^T\right]$

③ solve $\det(S - \lambda I) = 0$ for eigen values

④ solve $(S - \lambda I)u = 0$ using max eigen values to get the eigen vector

⑤ Take the eigen vector u_1 with largest $\lambda_1 \rightarrow$ this the 1st principle component.

Finding optimum number of Principle Components.

let's n number of features then there will be n eigen values and components.

a single λ tells how much this single component can explain the variance of original data.

eigen vector

explained_variance_ratio tell how percentage on component contribute to get the max variance.

when PCA does not work

data distribution circle