# Principal Component Analysis: PCA

Aims to transform the original features of a dataset into a new set of Uncorrelated variables called Principal Components, with the objective of retaining most important information while discarding the least important.

Objective :

Retain most important information

Maximize Variance And Reduce Redundancy

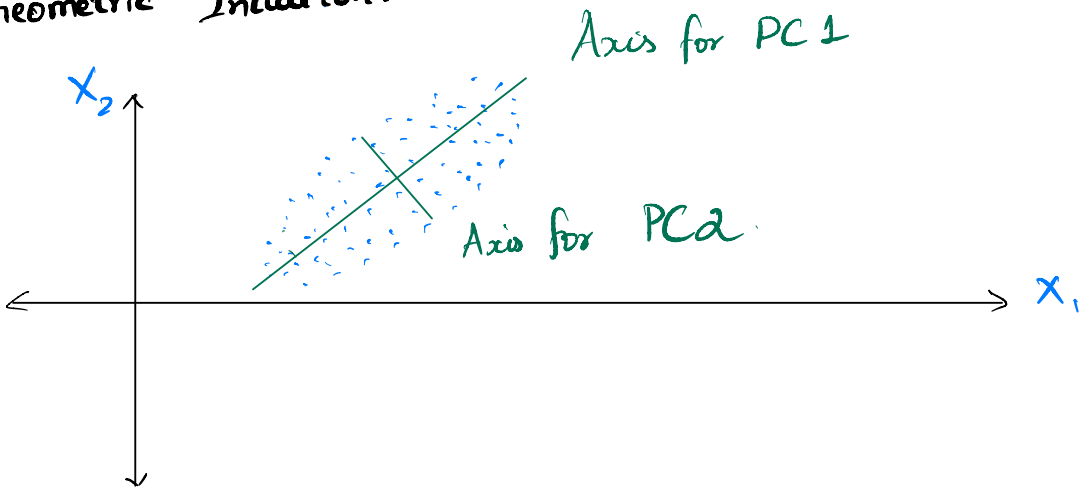
Achieved by Ensuring the principal components are orthogonal to each other.

Achieved by Finding the direction of Principal components in which it can capture the most variance.

Minimize Reconstruction Error

Achieved by Finding the principal component axis passing through the data mean and projecting the data points perpendicular to it.

## Geometric Intuition:




Dataset of $d$ features can be transformed into $k$ principal components, where $k \leq d$

Principal Components are ordered based on the amount of variance (information) they capture in the data.

Direction of PC 1 → Most variance

⋮

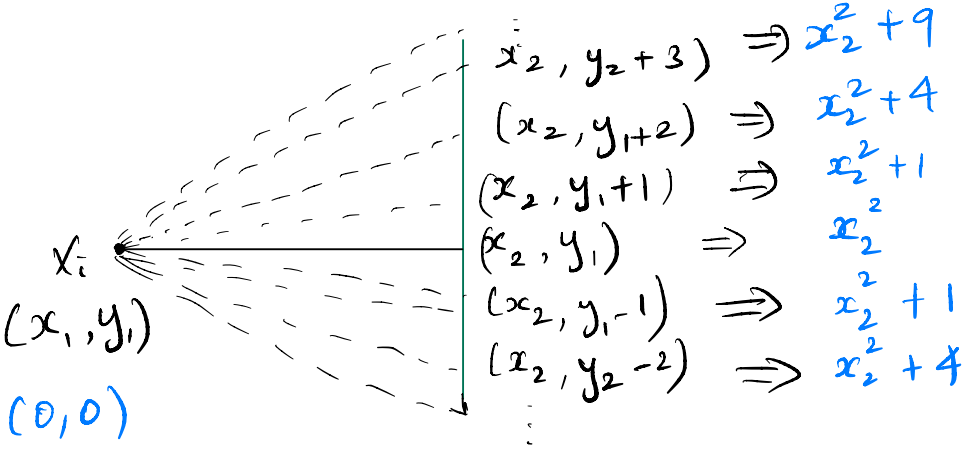
Direction of PC $d$ → Least variance.

Principal components are orthogonal to each other ensuring each PC captures unique and non-redundant aspect of the variability of the data.

Let's first solve how we are going to project our data points to these new principal component axes.

Assume we found our new axis

We need to project original data points onto this new line with minimal SSE




Triangle analogy: The straight "perpendicular" path is like one side of the triangle. It's always shorter than any other path you could take that forms a slanted side of the triangle

We need to find

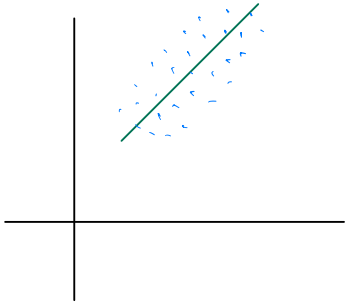
1) Direction of each principal component
2) New co-ordinates [w.r.t < new axis / original axis]

We can find these by solving our objectives.

② New co-ordinates:

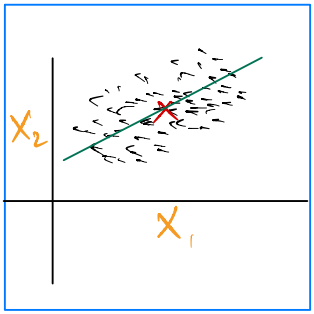
Assume we have the optimal direction for the principal component 1.

Let $\vec{e}$ be the best direction in which we can capture most variance.

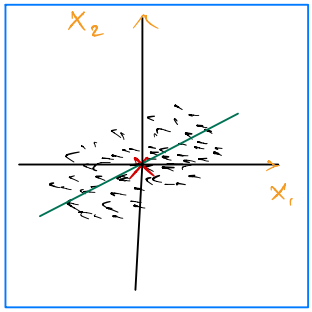


In order to find the new co-ordinates, we need a reference point, like the origin in our current co-ordinate system and then $\vec{e}$ will be our new basis.

To minimize SSE, the new axis line should pass through the data mean.

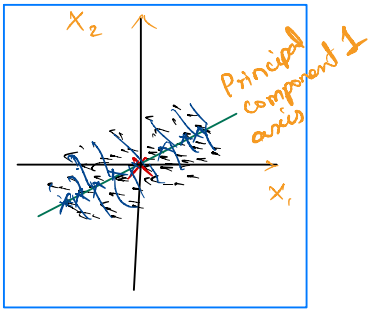



$X$
on original
co-ordinate system

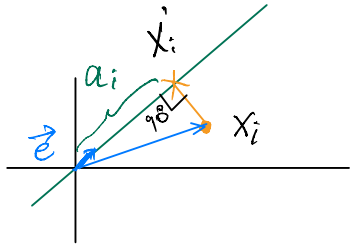



$X_{mean\_centered} = X - X_{mean}$
on original
co-ordinate system

Optimal line that minimizes the SSE & maximize the variance always go through the mean. So mean is the reference point or origin in new co-ord system. Mean center it to make MATH EASY




Project $X$ on new axis




$X_i$ - Co-ordinates of data point of interest in original feature space.

$X_i'$ - Co-ordinates of projected data point in original feature space

$a_i$ - New co-ordinate in the principal component with the basis $\vec{e}$

---

To find $a_i$ :

We know that angle b/w $(\overrightarrow{X_i - X_i'})$ & $\vec{e}$ is 90°.

So, dot product is 0 $\Rightarrow (X_i - X_i').e = 0$

Since, $X_i' = a_i.e$,

$(X_i - a_i.e).e = 0$

$X_i.e - a_i = 0$

$a_i = X_i.e$

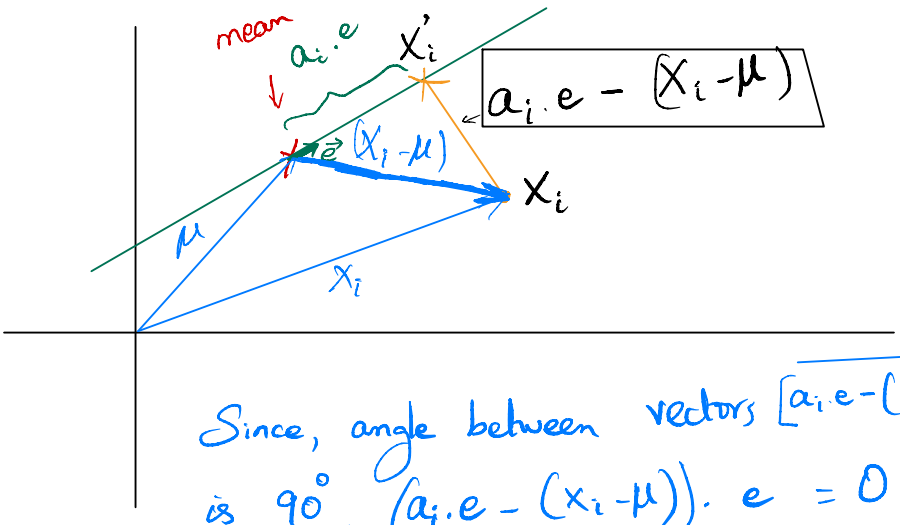
$a_i$ gives the magnitude of $X_i'$

If not mean centered, mean is our reference point.

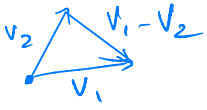
$a_i = (X_i - \mu).e$

$X_i' = \mu + a_i e$

If not mean centered




Reminder




Since, angle between vectors $\overrightarrow{[a_i e - (X_i - \mu)]}$ & $\vec{e}$ is $90^\circ$,

$$(a_i \cdot e - (X_i - \mu)) \cdot e = 0$$

$$a_i - (X_i - \mu) \cdot e = 0$$

$$a_i = (X_i - \mu) \cdot e$$

And, $X_i' = \mu + a_i \cdot e$

---

Two objectives: 1) Maximize Variance

2) Minimize SSE

(2) MINIMIZE SSE

$$SSE = \sum_{i=1}^{n} \| (\mu + a_i e) - X_i \|^2 = \sum_{i=1}^{n} \| a_i e - (X_i - \mu) \|^2$$

$$= \sum_{i=1}^{n} a_i^2 \underbrace{\|e\|^2}_{1} - 2 \sum_{i=1}^{n} a_i e^T (X_i - \mu) + \underbrace{\sum_{i=1}^{n} \|X_i - \mu\|^2}_{\text{Constant}}$$

Minimize SSE
$a_i$

$$SSE = \sum_i a_i^2 - 2\sum_i a_i e^T(x_i - \mu) + \sum_i \|x_i - \mu\|^2$$

$$\frac{\partial SSE}{\partial a_i} = 2a_i - 2e^T(x_i - \mu) \overset{\text{want}}{=} 0$$

$$a_i - e^T(x_i - \mu) = 0$$

$$a_i = e^T(x_i - \mu)$$

[Same result as the one we got when solved for the line passes through the MEAN]

## (1) MAXIMIZE VARIANCE AND FIND $\vec{e}$

Two ways to find $\vec{e}$ → By minimizing SSE
→ By Maximizing Variance

(i) MINIMIZE SSE

$$SSE = \sum_i a_i^2 - 2\sum_i a_i e^T(x_i - \mu) + \sum_i \|x_i - \mu\|^2$$

$$= \sum_i a_i^2 - 2\sum_i a_i^2 + \sum_i \|x_i - \mu\|^2$$

$$= -\sum_i a_i^2 + \sum_i \|x_i - \mu\|^2$$

$$= -\sum_i (e^T(x_i - \mu))^2 + \sum_i \|x_i - \mu\|^2$$

$$SSE = -\sum_i e^T (x_i - \mu)(x_i - \mu)^T e + \text{constant}(e)$$

$$= -e^T \left[ \sum_i (x_i - \mu)(x_i - \mu)^T \right] e + \text{ignore}$$

$$= -e^T \underbrace{\Sigma}_{\text{covariance Matrix}} e + \text{ignore constant}(e)$$

& it is constant w.r.t $e$

To minimize SSE, we have to minimize $-e^T \Sigma e$

or maximize $e^T \Sigma e$

(ii) MAXIMIZE VARIANCE

$$\text{Variance}(X_i' \text{ on } e) = E\left[ \left( \mu + a_i e - E[\mu + a_i e] \right)^2 \right]$$

$$= E\left[ (\mu + a_i e - \mu)^2 \right]$$

$$= E\left[ (a_i e)^2 \right]$$

$$= E\left[ e^T (x_i - \mu) \cdot e^T (x_i - \mu) \right]$$

$$= E\left[ e^T (x_i - \mu)(x_i - \mu)^T e \right]$$

$$= e^T E\left[ (x_i - \mu)(x_i - \mu)^T \right] e$$

Variance ($X_i$ on $e$) $= e^T \Sigma e$

Again, to maximize variance, we have to

maximize $e^T \Sigma e$

[Same as minimizing SSE ; it tells us both are same objectives]

---

## Yet Another Intuition:

After projecting the data points $X$ onto a principal component of direction vector $\vec{e}$, the new data points vector is given as $e^T X$

we know, on point level $a_i = X_i e$

So, lets revisit two things Variance & Magnitude.

| | |
|---|---|
| $\sigma^2 = \frac{1}{n-1} \sum_i (V_i - \bar{V})^2$ | $\|V\|^2 = \sum_i V_i^2$ |

Variance and Vector magnitude are montonically related when mean-centered ; so maximizing one maximizes the other.

Hence, Maximizing variance is equal to maximizing the vector magnitude $\|e^T X\|^2$ [note: $\|e\| = 1$]

$$\|e^T X\|^2 = (e^T X)(e^T X)^T$$

$$= e^T X X^T e$$

$$= e^T \Sigma e$$

Once again, it is clear that we need to maximize $e^T \Sigma e$ in order to maximize variance and minimize SSE.

$e^T \Sigma e$ is the quadratic form

Let's solve $\vec{e}$ that maximizes $e^T \Sigma e$

This is a constrained optimization. $\Rightarrow \|e\| = 1$

$\Rightarrow e^T e = 1$

We can use Lagrange Multiplier to solve this constrained optimization.

Lagrange Multiplier:

$$\text{MAX } L = \underbrace{e^T \Sigma e}_{\text{objective}} - \lambda \underbrace{(e^T e - 1)}_{\text{Constraint}}$$

↳ Lagrangian multiplier = ratio of the tangents at solution point.

idea: $\partial\, e^T \Sigma e$ & $\partial\,(e^T e - 1)$ have same tangent direction at the solution

$$\frac{\partial L}{\partial e} = \frac{\partial\, e^T \Sigma e}{\partial e} - \lambda \frac{\partial\, e^T e}{\partial e}$$

$$= 2\Sigma e - 2\lambda e \overset{\text{want}}{=} 0$$

$$\Rightarrow \underbrace{\Sigma e}_{\substack{\downarrow \\ \text{Covariance Matrix} \\ D \times D}} = \underbrace{\lambda e}_{\substack{\downarrow \\ \text{Scalar}}}$$

"e is a special vector for $\Sigma$ So, $\Sigma$ does not rotate e by multiplication"

① $\Sigma$ is symmetric

② $\Sigma = X^T X \Rightarrow$ Positive Definite

$\Rightarrow \forall$ vector $z$, $z^T \Sigma z \geq 0$

③ Eigen vectors of $\Sigma$ : $e_1, e_2, \ldots e_D$ are orthonormal : $e_i^T e_i = \|e_i\| = 1$ ; $e_i^T e_j = 0$ (perpendicular)

Spectral decomposition of $\Sigma$

$$\Sigma = \underbrace{\begin{bmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_D \\ | & | & & | \end{bmatrix}}_{E} \cdot \underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_D \end{bmatrix}}_{\text{Diag}} \cdot \underbrace{\begin{bmatrix} - & e_1 & - \\ - & e_2 & - \\ & \vdots & \\ - & e_D & - \end{bmatrix}}_{E^T}$$

$D \times D$

diag - eigen vals sorted

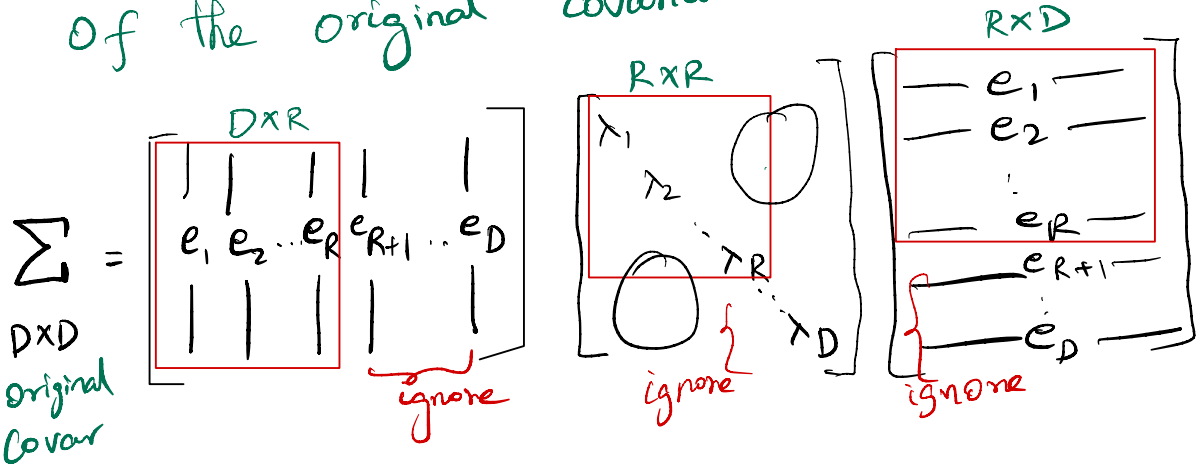
$\lambda_1, \lambda_2, \ldots \lambda_D \geq 0$

# Dimensionality Reduction:

want: $R < D$ dimensions for projections.

$\Rightarrow$ Project $X$ onto $e_1\ e_2 \dots e_R$ $\Rightarrow$ R co-ordinate per data point

Covariance Matrix of new data (principal components is R-dimensional APPROXIMATION of the original covariance matrix.




$\Sigma_{new}$ R-dim data DXD

$$\begin{bmatrix} | & | & & | \\ e_1 & e_2 & \dots & e_R \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_R \end{bmatrix} \begin{bmatrix} - & e_1 & - \\ - & e_2 & - \\ & \vdots & \\ - & e_R & - \end{bmatrix}$$