- PCA (A feature extracⁿ technique)

  Unsuper ml problem

  Complex technique

  old technique

  reliable

  Underlying math is difficult.

- Aim: To reduce curse of dimensionality

- PCA converts higher dimensional dataset to lower dimen.. while keeping essense of data.

- Why PCA is used

  benefits:

  HD data to LD data.

  1) Faster execution of algorithm.

  2) Visualization : HD data we cannot visualized
     LD data can be visualized.

- Geometric intution

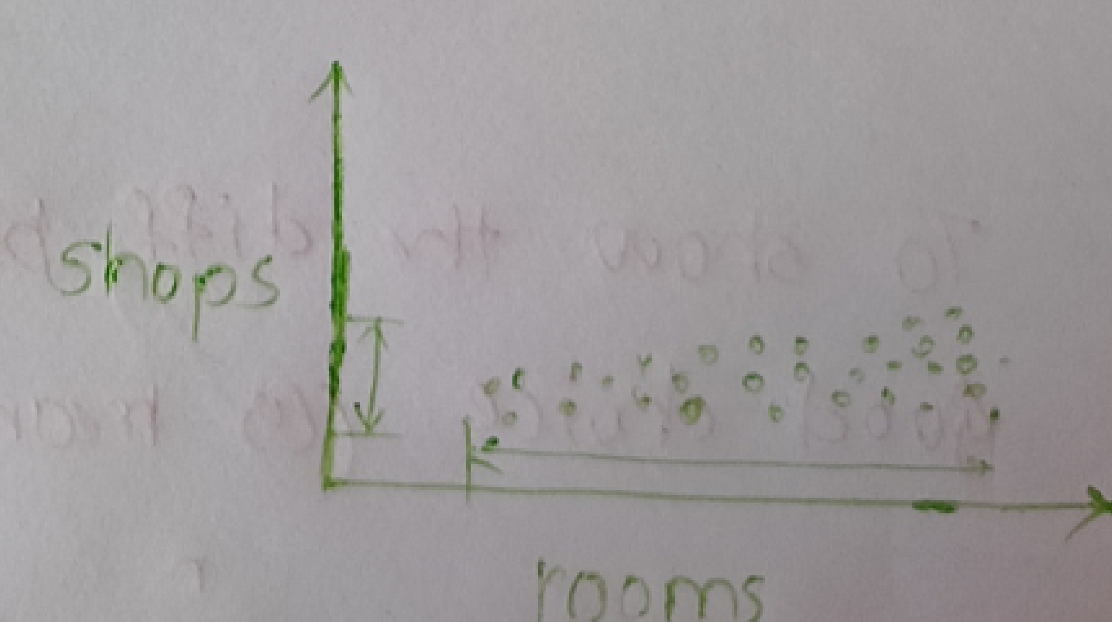
Suppose we want to predict house price with features no. of rooms & no. of shops near it.

(no. of rooms is feature strong) > (shops weak feature)

↑

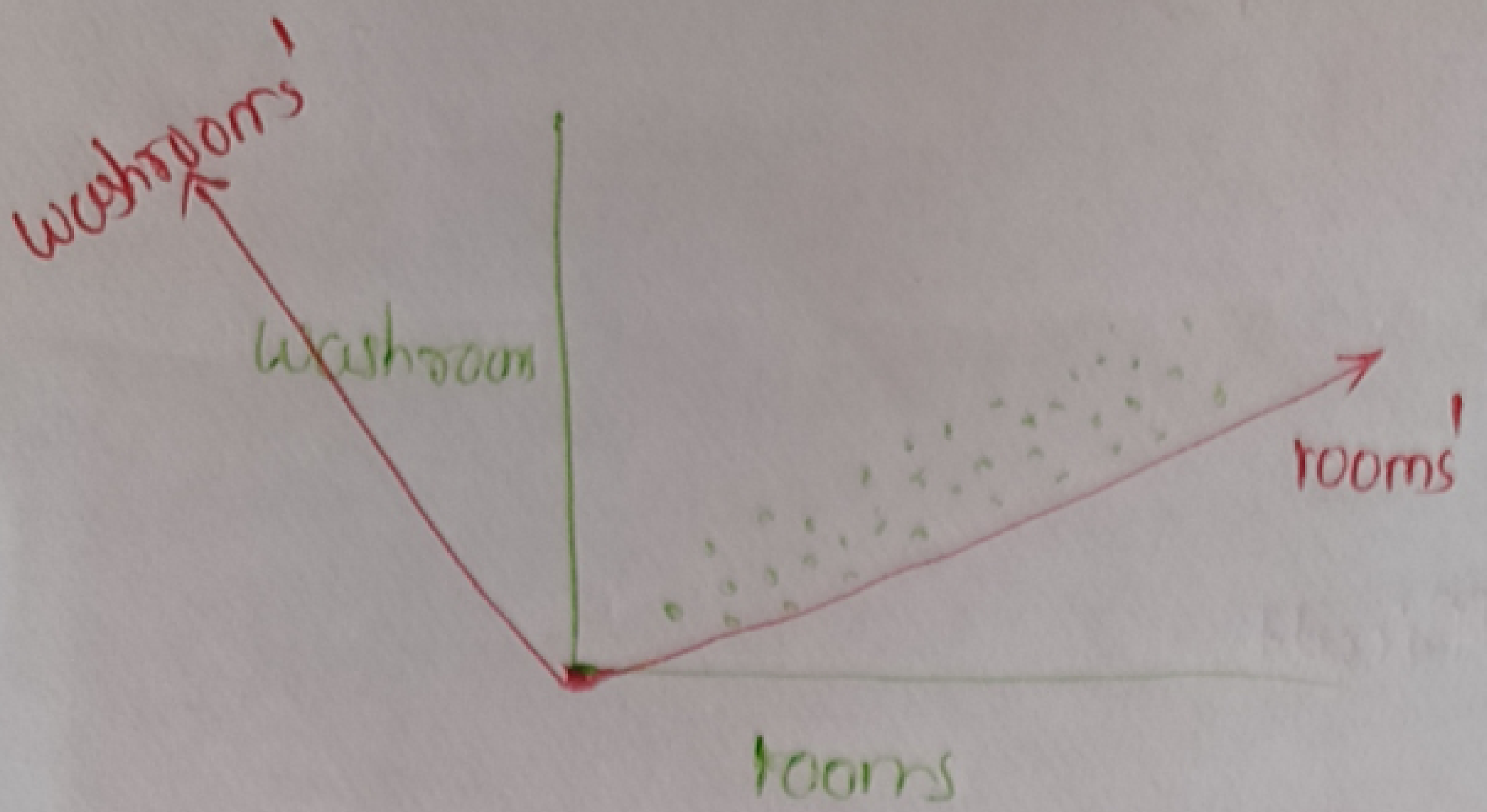
We will keep this

We keep data axis with max variance. hence rooms.



Pto.

but if features are rooms & washrooms



Linear relation
difficult to eliminate
one

Will rotate the dimensions
such that,
Variance among one of the
axis is more.
We will keep that new
axis which holds essense of both the previous axeses
Those new axeses are called principle components P.C.

$$P.C. \le n \; ; \; n \Rightarrow \text{actual dims}$$

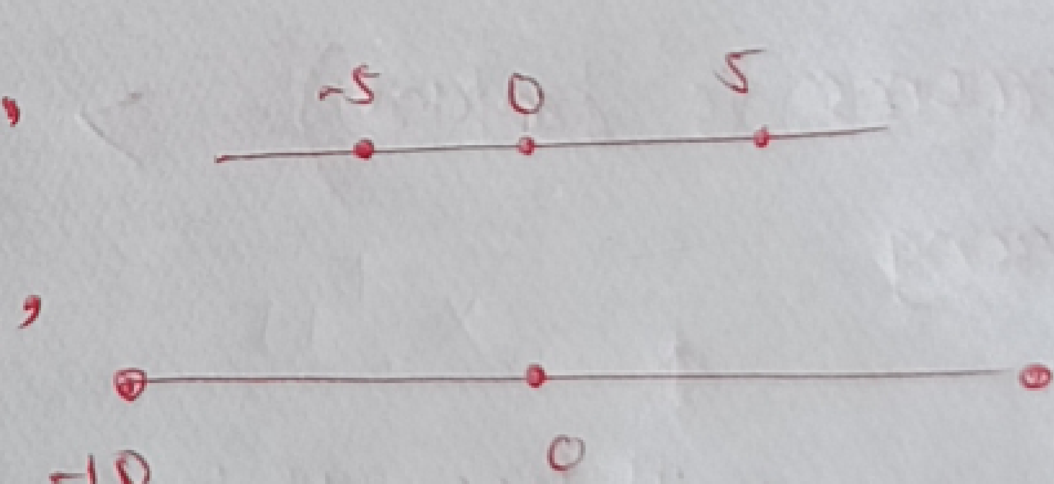
- Why variance is so imp?

What is variance?

↳ gives info about spread of data.

eg. [-5, 0, 5] → mean '0'

[-10, 0, 10] → mean '0'



To show the diff. bet. these two datasets, Variance is good choice. (as mean gives central tendency)

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

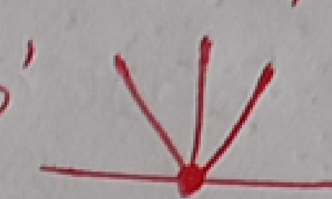
$(-5, 0, 5) \Rightarrow (25+0+25)/3 = 50/3$

$(-10, 0, 10) \Rightarrow (100+0+100)/3 = 100/3$ $\therefore \frac{100}{3} > 50/3$

mind it !

~~spread is not~~ variance, is not spread

~~spread~~ is proportional to spread
var

mean absolut deviation (MDA) gives exact spread.
But MDA is not differentiable at '0'  . Hence not used in PCA

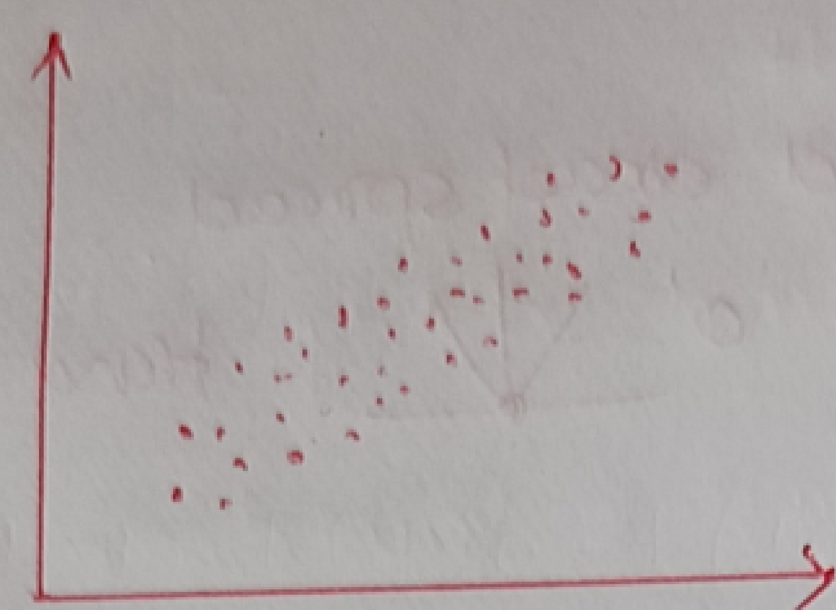
why var is so imp for PCA?

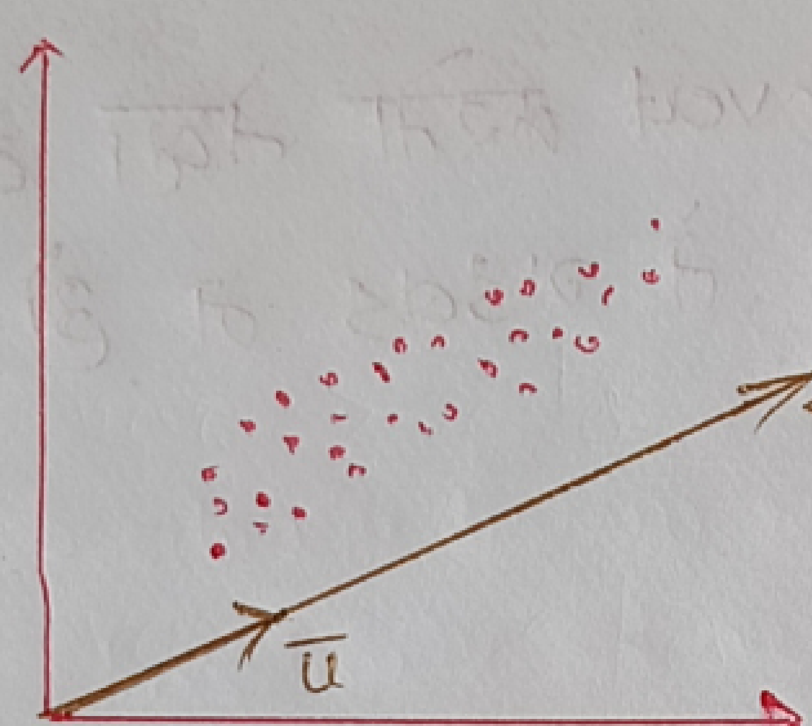
जेव्हा HD to LD मध्ये आपण convert करतो तेव्हा data pts जो परस्परसंबंधी relation आहे ते गडबड न हो म्हणून Var. max. करावे लागते.

Mathematics of PCA.

* Problem formulation:

आपल्याला var maximize करायचा आहे.





Project of any pt. on new axis $= \dfrac{\bar{u} \cdot \bar{x}_i}{|u|}$

$= \bar{u} \cdot \bar{x}_i \quad ; |u| = 1$

$= u^T \cdot x_i$

$= [x_1 \cdot y_1] \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$

$= x_1 x_2 + y_1 y_2$

projection will be a scalar

So, $var = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_{mean})^2$

येथे x_i च projection : $u^T . x_i$

साऱ्या pts च्या mean pt. च projection : $u^T . \bar{x}$

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

आता आपल्या PCA साठी असा unit vector find करायचा, ज्याची 'var' value maximum असणार.

to understand the solution,

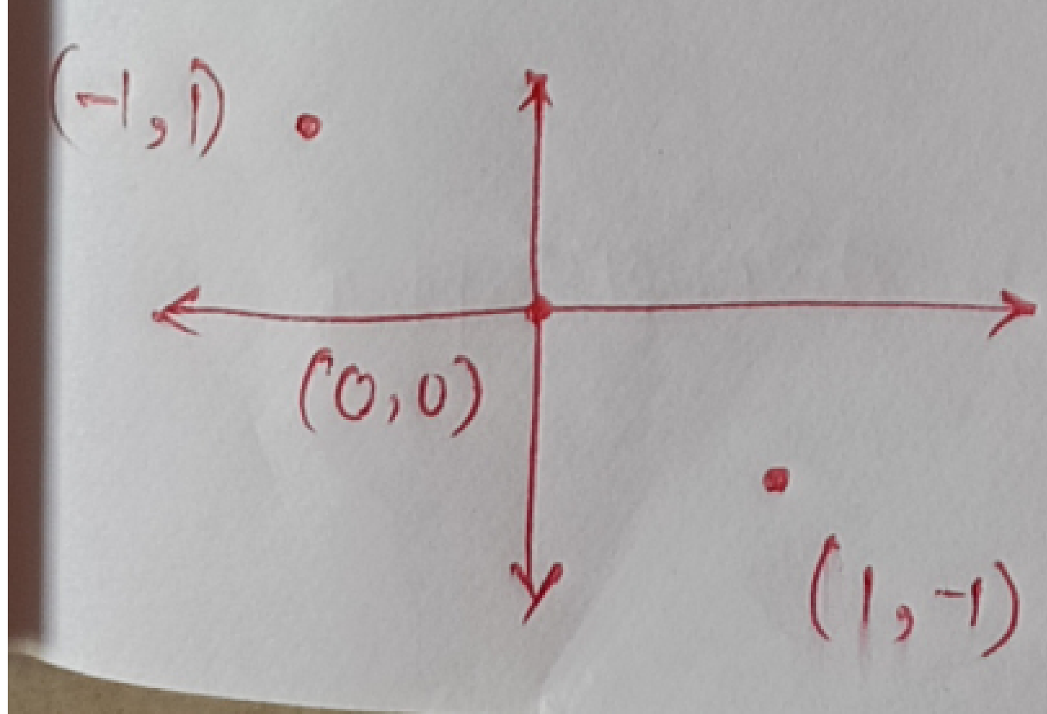
u have to know covariance and eigen values, vectors.

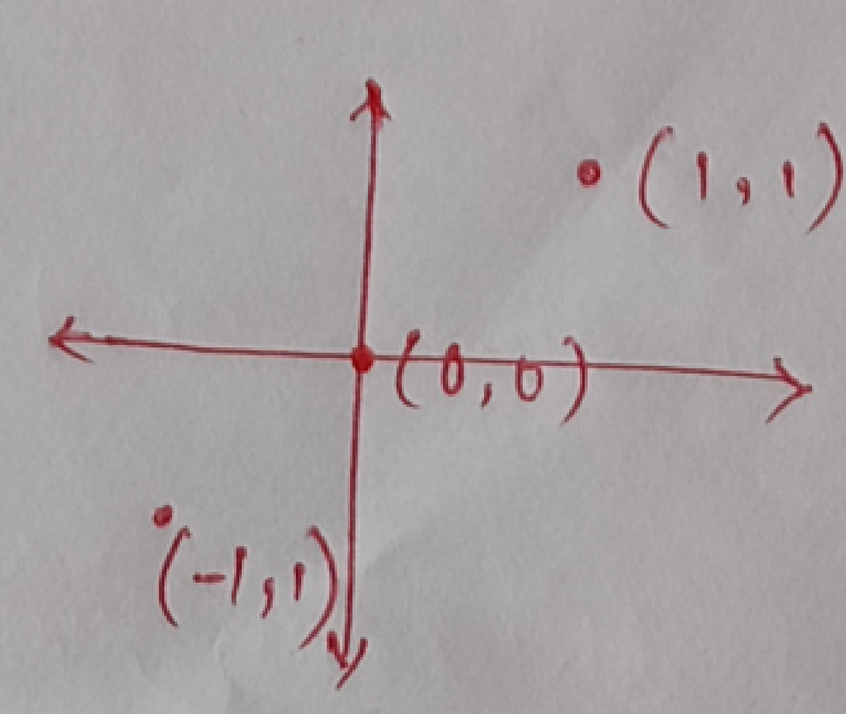
mean = $\cancel{(-10+0+10)}$

$\frac{(-10+0+10)}{3} = 0$
$\frac{(-5+0+5)}{3} = 0$
} mean फक्त central tendency दर्शविते, spread बद्दल सांगत नाही

var $\Rightarrow$ $\frac{(-10)^2+0+(10)^2}{3} = \frac{100}{3}$
$\frac{(-5)^2+0+(5)^2}{3} = \frac{50}{3}$
} var spread दर्शविते

पण, var axises ची relationship दाखवत नाही,
for eg.





या दोन्ही साठी var same आहे, म्हणजे येथे var fail करतं.

Covariance, data मध्ये x आणि y मध्ये काय relationship असेल ते दर्शवत.

$$Covar = \frac{(x_1 \times y_1) + (x_2 \times y_2) + \cdots (x_n \times y_n)}{n}$$

∴ for prev. eg.

$$\frac{(-1\times1)+(0\times0)+(1\times(-1))}{3} \neq \frac{(-1\times(-1))+0+(1\times1)}{2}$$

$$-\frac{2}{3} \neq \frac{2}{3}$$

+ve co-var ⇒ x ↑es and y also ↑es

−ve co-var ⇒ x ↓es & y ↑es

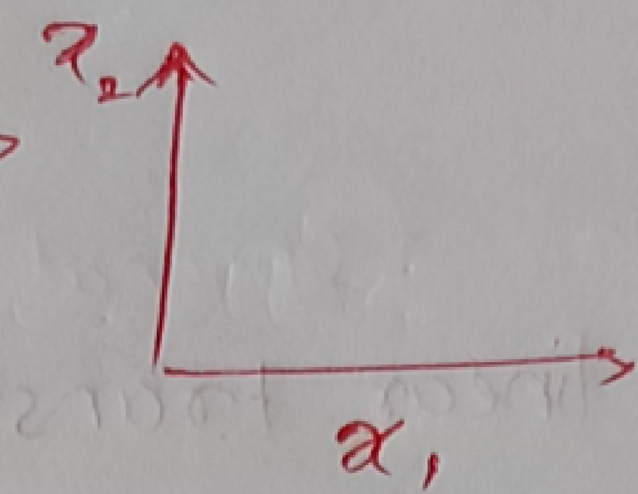
Covar and co-relation are same just आपण co-relation ला (−1 to 1) च्या range मध्ये देवतो.

Covar metrix

$\therefore$

x_1	x_2

$\longrightarrow$



$\longrightarrow$ $\begin{array}{c|cc} & x_1 & x_2 \\ \hline x_1 & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 2\times 2 \end{bmatrix}$}} \\ x_2 & & \end{array}$

$$\begin{array}{c} x_1 \\ x_2 \end{array}\begin{bmatrix} covar(x_1, x_1) & covar(x_2 x_1) \\ covar(x_1, x_2) & covar(x_2, x_2) \end{bmatrix}$$
$$\qquad x_1 \qquad\qquad\qquad x_2$$

$$\therefore \begin{array}{c} x_1 \\ x_2 \end{array}\begin{bmatrix} var(x_1) & cov(x_2, x_1) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix} \quad \cdots \quad cov(x_i, x_i) = var(x_i)$$
$$\qquad x_1 \qquad\qquad\qquad x_2$$

$$\begin{array}{c} x_1 \\ x_2 \end{array}\begin{bmatrix} var(x_1) & cov(x_2 x_1) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix} \quad \cdots \quad cov(x_i x_j) = cov(x_j x_i)$$
$$\qquad x_1 \qquad\qquad\qquad x_2$$

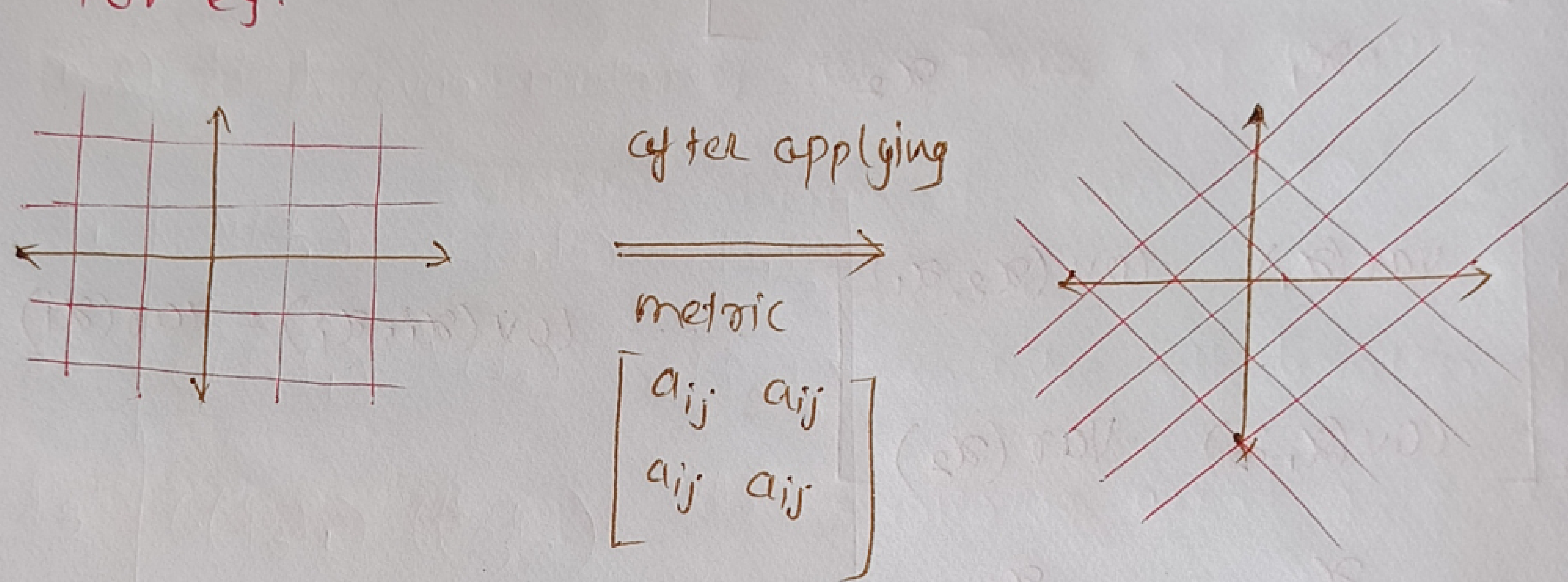
Hence in cov metrix, diagonal elements are var of that clmn, and remaining metrix is symmetric.

Linear transformations, eigen values, eigen vectors, metrices.

What are metrices,

Metrices are basically linear transformations, ज्यांना जेव्हा आपण एखाद्या co-ordinate system वर apply करू तर त्या co-ordinate system चे सारे points or vectors transform होतात.

for eg.



What are eigen values and eigen vectors?

Eigen vectors?

these are speical vectors, ज्यांच्यावर transformation apply केल्यावर direction change नाही होत. Magnitude may change.

एखादं transformation apply केल्यावर in 2D, 2 ~~minimum~~ eigen vectors मिळतात

explore the website 'Geogebra'

Eigen values?

Eigen vector किती stretch or shrink करतोय after transformation.
i.e. किती times stretch/shrink होतोय.

In eigenvectors it is true.

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

A → (matrix) or linear transform
$\overline{V}$ → eigen vector
λ → eigen value
$\overline{V}$ → eigen vector.

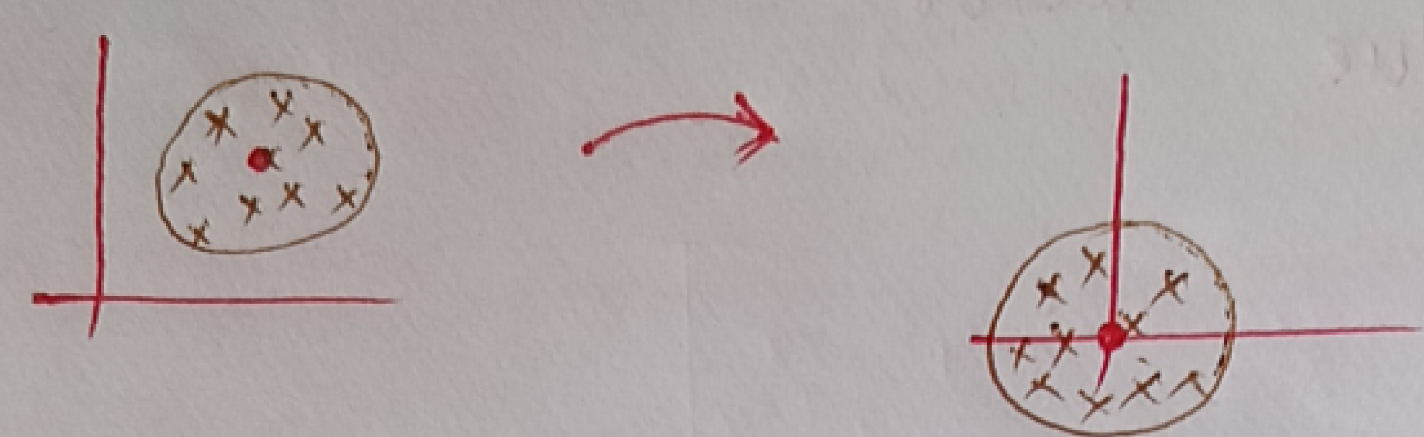
Eigendecomposition of Covariance matrix.

Largest eigenvector of covar. matrix always points into the direction of largest variance of the data, and magnitude of this vector equals to the corresponding eigen value.

थोडक्यात आपल्याला covar matrix चं decomposition करायचयं ~~आणि जो सगळ्यात~~ म्हणजे eigen values आणि eigen vectors find करायचेत, आणि जो सगळ्यात मोठा eigen vector आहे (ज्याची eigen value सगळ्यात जास्त आहे) त्या vector साठीच data चा variance सगळ्यात जास्त येणार.

Step by step solution of PCA.

Step 1) Make data mean centric (optional step) (to substract mean of data from each individual point) (समूह के data mean or center पे ले जाना)



Step 2) Find covar. matrix

$$\begin{array}{c|ccc} & f_1 & f_2 & f_3 \\ \hline f_1 & V(f_1) & C(f_1 f_2) & C(f_1 f_3) \\ f_2 & C(f_2 f_1) & V(f_2) & C(f_2 f_3) \\ f_3 & C(f_3 f_1) & C(f_3 f_2) & V(f_3) \end{array}$$

Step 3) Find eigen vector and eigen values.

$$\lambda_1 \rightarrow PC_1 \qquad \lambda_2 \rightarrow PC_2 \qquad \lambda_3 \rightarrow PC_3$$