## (1) Where should i use PCA??

*1) if we want to reduce no of variables but not able to identify variable to completly remove from consideration.

2) if i want to ensure my variables are independent

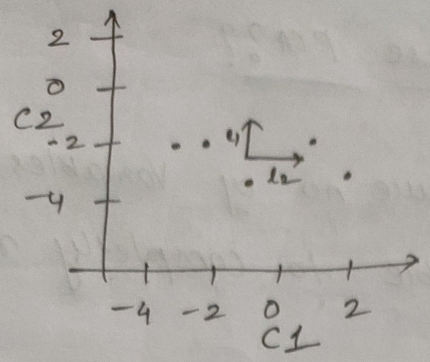
3) if i am comfortable of making my variable less interpretable

## (2) How PCA works ??

calculate a matrix which tells how variables relate each other

then break the matrix into ~~two~~ 2 components

1. Direction
2. Magnitude



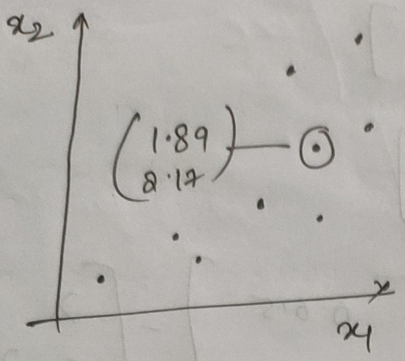

Which line will be the best?

Lets dive into maths behind this

Lets take a sample dataset of $n=10$.

| $x_1$ | $x_2$ |
|---|---|
| 2.0 | 2.4 |
| 0.5 | 0.7 |
| 2.1 | 2.9 |
| 2.6 | 2.2 |
| 2.8 | 3.0 |
| 2.9 | 1.0 |
| 1.0 | 1.6 |
| 1.5 | 1.9 |
| 1.2 | 2.1 |
| 2.3 | 3.9 |
| Mean = 1.89 | Mean 2.17 ($\bar{x_2}$) |

If we plot all data points




step 1:-

Subtract mean from corresponding data to recentre

2) Draw the scatter plot to view

$$X_1 = \bar{x}_1 - x_1 \qquad X_2 = \bar{x}_2 - x_2$$

3) Now we get a matrix of 10×2 which will have <u>mean zero</u>

step 2:-

Calculate the covariance of the new dataset.

$$C = \frac{1}{N-1}\begin{pmatrix} (x,x) & (x,y) \\ (y,x) & (y,y) \end{pmatrix}$$

$$C = \begin{bmatrix} 0.616 & 0.615 \\ 0.615 & 0.716 \end{bmatrix}$$ 2×2 matrix

compute eigen values.

$$\lambda I = \lambda \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$C - \lambda I = \begin{bmatrix} 0.616 & 0.615 \\ 0.615 & 0.716 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 0.616 - \lambda & 0.615 - 0 \\ 0.615 - 0 & 0.6716 - \lambda \end{bmatrix}$$

$$= \det \begin{bmatrix} 0.616 - \lambda & 0.615 \\ 0.615 & 0.716 - \lambda \end{bmatrix}$$

$$= \left[ (0.616 - \lambda)(0.716 - \lambda) - (0.615)^2 \right]$$

$$= 0.441 - 0.661\lambda - 0.716\lambda + \lambda^2 - 0.37$$

$$= \lambda^2 - 1.33\lambda + 0.0631$$

eigen values $\therefore$ $\boxed{\lambda_1 = 1.28 \quad \lambda_2 = 0.0492}$

So $\begin{bmatrix} 0.616 - 1.28 & 0.615 \\ 0.615 & 0.716 - 1.28 \end{bmatrix}$

$$B = \begin{bmatrix} -0.66 & 0.615 \\ 0.615 & -0.56 \end{bmatrix}$$

$$B\bar{x} = 0$$

$$B \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving the ~~ma~~ Above matrix we get a set of eigen vector

$$x_1 \begin{Bmatrix} x_1 = 0.678 \\ x_2 = 0.735 \end{Bmatrix}$$

~~for~~ similarly for $\lambda_2 = \begin{Bmatrix} 0.73 \\ -0.67 \end{Bmatrix}$

** 1. total Variance calculation :-

$\lambda_1 = 1.28$

$$\frac{1.28}{0.67 + 0.73} = 96\%$$

PC1

$\lambda_2 = 0.049$

$$\frac{0.049}{0.73 + 0.67} = 3.7$$

PC2

Now eigen vector :-

$$\lambda_1 = \begin{pmatrix} 0.67 \\ 0.73 \end{pmatrix} \qquad \lambda_2 = \begin{pmatrix} 0.73 \\ -0.67 \end{pmatrix}$$

* eigen vector 1, 0.678 move right direction
& 0.735 direction are up.

* eigen vec 2, 0.73 moves right & -0.67
direction up

Hence the eigen vector with highest eigen value is the principal component.

Summary :-

1) convert n-dimension data to D-dimension
2) To find direction where spread is wide
3) Finding covariance
4) compute eigen values i.e $\lambda_1 > \lambda_2 > \lambda_3 \cdots \lambda_d$.
5) compute eigen vectors i.e $V_1 \perp V_2 \perp^{or} V_3 \cdots \perp^{or} V_d$
6) use of $\lambda$ :- for maximum variance

$\lambda_1 = 3$ $\quad \frac{\lambda_1}{\lambda_1+\lambda_2} = 60\%$ Variance achieved

$\lambda_2 = 2$