PCA #example

X	Y	PCA_1	PCA_2
x_1 4	y_1 11	? P_{11}	? P_{21}
x_2 8	y_2 4	? P_{12}	? P_{22}
x_3 13	y_3 5	? P_{13}	? P_{23}
x_4 7	y_4 14	? P_{14}	? P_{24}

Step 1 : Mean of variable:

Sample = N = 4
Features = n = 2

$$\bar{x} = \frac{4+8+13+7}{4} = 8$$

$$\bar{y} = \frac{\cancel{4+8+13} \; 11+4+5+14}{4} = 8.5$$

Step #02 : Covariance Matrix (x,y)

$(x,y) = (x,x) \quad (x,y) \quad (y,x) \quad (y,y)$

$$\begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix} \rightarrow \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$cov(x,x) = \frac{1}{N-1}\left[(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + (x_4-\bar{x})^2\right]$$

$$= \frac{1}{4-1}\left[16+0+25+1\right] \Rightarrow \frac{1}{3}(42) = 14$$

~~[illegible]~~

$$cov(x,y) = \frac{1}{N-1}\left[(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + (x_3-\bar{x})(y_3-\bar{y}) + (x_4-\bar{x})(y_4-\bar{y})\right]$$

$$= \frac{1}{4-1}\left[(4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)\right]$$

$$= \cancel{\frac{1}{3}\left[+10 + (-45) + (25) + (56)\right]}$$

$$= \cancel{72.3}\ \frac{1}{3}\left[-10 + 0 - 17.5 + (-5.5)\right]$$

$$= \frac{1}{3}(-33) \Rightarrow -11$$

$$cov(x,y) = cov(y,x)$$

$$cov(y,y) = \frac{1}{N-1}\left[(11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2\right]$$

$$= \frac{1}{3}\left[6.25 + 20.25 + 12.25 + 30.25\right]$$

$$= \frac{1}{3}(69) \Rightarrow 23$$

put in ①

$$\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Scanned with CamScanner

Step #03:

Eigen Values:

$\det(S - \lambda I)$

$$\left| \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right|$$

$$\left| \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \right|$$

$$= (14-\lambda)(23-\lambda) - (-11)(-11)$$

$$= 322 - 14\lambda - 23\lambda + \lambda^2 - 121$$

$$= \lambda^2 - 37\lambda + 201$$

Apply quadratic formula:

$$= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{+37 \pm \sqrt{(-37)^2 - 4(1)(201)}}{2(1)}$$

$$= \frac{37 \pm \sqrt{1369 - 804}}{2}$$

$$= \frac{37 \pm \sqrt{565}}{2}$$

$$= \frac{37 + \sqrt{565}}{2}, \quad = \frac{37 - \sqrt{565}}{2}$$

$$= 30.38, \quad = 6.615$$

$$\lambda_1 = 30.38 \qquad \lambda_2 = 6.615$$

As,

$\lambda_1 > \lambda_2$

$\lambda_1 = 1^{ST}$ PCA

Step # 04:

Eigen vector:

$$(S - \lambda I)u_1 = 0$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$(14-\lambda)u_1 - 11u_2 = 0 \rightarrow \text{(ii)}$$

$$-11u_1 - (23-\lambda)u_2 = 0$$

$$(14-\lambda)u_1 = 11u_2 \quad \text{from (ii)}$$

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda}$$

Let this equal to $t = 1$:

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda} = t$$

$$\frac{u_1}{11} = t \quad , \quad \frac{u_2}{14-\lambda} = t$$

$$\boxed{u_1 = 11} \quad \boxed{u_2 = 14-\lambda}$$

Eigen vector for λ_1:

$$e_1 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix} = \begin{bmatrix} 11 \\ 14-30.38 \end{bmatrix} = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix}$$

Eigen vector for λ_2:

$$e_2 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 14-\lambda_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 14-6.615 \end{bmatrix} = \begin{bmatrix} 11 \\ 7.39 \end{bmatrix}$$

Step # 05

<u>Normalizing:</u>

$$e_1 = \begin{bmatrix} \dfrac{11}{\sqrt{(11)^2 + (-16.38)^2}} \\ \dfrac{-16.38}{\sqrt{(11)^2 + (-16.38)^2}} \end{bmatrix} = \begin{bmatrix} 0.5575 \\ -0.8301 \end{bmatrix}$$

$$e_2 = \begin{bmatrix} \dfrac{11}{\sqrt{(11)^2 + (7.39)^2}} \\ \dfrac{7.39}{\sqrt{(11)^2 + (7.39)^2}} \end{bmatrix} = \begin{bmatrix} 0.8301 \\ 0.5577 \end{bmatrix}$$

Step #06

Derive new dataset: PCA_1 & PCA_2

<u>for PCA_1:</u>

$$P_{11} = e_1^t \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 \cdot -0.8301 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$P_{11} = \begin{bmatrix} -2.23 \cdot 2.07525 \end{bmatrix}$$

$\boxed{\cancel{P_{11} = 0.15479}}$ $P_{11} = \boxed{-4.30525}$

$$P_{12} = e_1^t \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 \cdot -0.8301 \end{bmatrix} \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$$

$\boxed{P_{12} = +3.73545}$

$$P_{13} = e_1^t \begin{bmatrix} 13-8 \\ 5-8.5 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.5575 & -0.8301 \end{bmatrix} \begin{bmatrix} 5 \\ -3.5 \end{bmatrix}$$

$$= \begin{bmatrix} 2.7875 + 2.90535 \end{bmatrix}$$

$$\boxed{P_{13} = 5.69285}$$

$$P_{14} = e_1^t \begin{bmatrix} 7-8 \\ 14-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 & -0.8301 \end{bmatrix} \begin{bmatrix} -1 \\ 5.5 \end{bmatrix}$$

$$P_{14} = \begin{bmatrix} -0.5575 - 4.5655 \end{bmatrix}$$

$$\boxed{P_{14} = -5.1230}$$

For $\underline{\underline{PCA_2}}$

$$P_{21} = e_2^t \begin{bmatrix} -4 \\ +2.5 \end{bmatrix}$$

$$= \begin{bmatrix} -4 \\ 2.5 \end{bmatrix} \begin{bmatrix} 0.8301 & 0.5577 \end{bmatrix}$$

$$= -3.3204 + 1.39425$$

$$\boxed{P_{21} = -1.92615}$$

$$P_{22} = e_2^t \begin{bmatrix} 0 \\ -4.5 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ -4.5 \end{bmatrix} \begin{bmatrix} 0.8301 & 0.5577 \end{bmatrix}$$

$$\boxed{P_{22} = -2.50965}$$

$$P_{23} = \begin{bmatrix} 5 \\ -3.5 \end{bmatrix} \begin{bmatrix} 0.8301 & 0.5577 \end{bmatrix}$$

$$= \begin{bmatrix} 4.1505 - 1.95195 \end{bmatrix}$$

$$\boxed{P_{23} = 2.19855}$$

$$P_{24} = \begin{bmatrix} -1 \\ 5.5 \end{bmatrix} \begin{bmatrix} 0.8301 & 0.5577 \end{bmatrix}$$

$$= \begin{bmatrix} -0.8301 + 3.06735 \end{bmatrix}$$

$$\boxed{P_{24} = 2.23725}$$

x_1	x_2	PCA1	PCA2
4	11	-4.30525	-1.92615
8	4	3.73545	-2.50965
13	5	5.69285	2.19855
7	14	-5.123	2.23725

PCA:

It is most commonly used unsupervised machine learning algorithm. It is a method of dimensionality reduction, feature extraction that transforms the data from "d-dimensional space" into a new co-ordinate system of dimension p, where p<=d

GOALS

- To reduce dimension of data set.
- To detect correlation b/w variables.
- To identify patterns.

* $\frac{1}{2}$ Advantages

~~[illegible]~~
~~[illegible]~~

- ↳ Decrease requirement for memory
- ↳ Lack of redundancy of data
- ↳ Small database representation.
- ↳ Reduce complexity.

* $\frac{1}{2}$ Disadvantages:

- ↳ The trade-off b/w information loss & dimensionality reduction.
- ↳ Low interpretability of principal components.