PCA - Example

Consider the dataset:-

Feature	Sample1	Sample2	Sample3	Sample 4
a	4	8	13	7
b	11	4	5	14

Step1:-

No. of features $n = 2$ (a, b)

No. of samples $N = 4$ (sample1, sample2, sample3, sample4)

Step2:-

Compute Mean:

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

$$\bar{b} = \frac{11+4+5+14}{4} = 8.5$$

Step3:-

Compute covariance matrix between features in the given dataset, ordered features are as:

(a, a), (a, b), (b, a), (b, b):

$$\mathrm{cov}(a,a) = \frac{1}{N-1}\sum_{K=1}^{N}(a_i-\bar{a})(a_i-\bar{a})$$

$$= \frac{1}{N-1}\sum_{K=1}^{N}(a_i-\bar{a})^2 \quad \text{[for same features]}$$

$$= \frac{1}{4-1}\left[(4-8)^2+(8-8)^2+(13-8)^2+(7-8)^2\right]$$

$$= \frac{4^2+0+5^2+1^2}{3} = \frac{16+0+25+1}{3} = 14$$

$\boxed{\mathrm{cov}(a,a)=14}$

$$cov(a,b) = \frac{1}{N-1} \sum_{K=1}^{N} (a_i - \bar{a})(b_i - \bar{b})$$

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

$$= \frac{1}{3} [-10 + 0 \cancel{+} - 17.5 - 5.5]$$

$$= -33/3 = -11$$

$$\boxed{cov(a,b) = -11}$$

$$cov(b,a) = \frac{1}{N-1} \sum_{K=1}^{N} (b_i - \bar{b})(a_i - \bar{a})$$

$$= cov(a,b) = -11$$

$$\boxed{cov(b,a) = -11}$$

$$cov(b,b) = \frac{1}{N-1} \sum_{K=1}^{N} (b_i - \bar{b})(b_i - \bar{b})$$

$$= \frac{1}{N-1} \sum_{K=1}^{N} (b_i - \bar{b})^2$$

$$= \frac{1}{4-1} [(11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2]$$

$$= \frac{69}{3} = 23$$

$$\boxed{cov(b,b) = 23}$$

covariance matrix can be written as :

$$S = \begin{bmatrix} cov(a,a) & cov(a,b) \\ cov(b,a) & cov(b,b) \end{bmatrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Step 4:-

Compute Eigen values & Eigen vectors

To compute Eigen values det $(S-\lambda I)=0$

$$I \text{ (Identity Matrix)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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

$$\det\left(\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = 0$$

$$\Rightarrow \det\left(\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \right) = 0$$

$$\Rightarrow (14-\lambda)(23-\lambda) - (-11 \times -11) = 0$$

$$\Rightarrow 322 - 14\lambda - 23\lambda + \lambda^2 - 121 = 0$$

$$\Rightarrow \lambda^2 - 37\lambda + 201 = 0$$

'λ' can be calculated by quadratic equation:

$$\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a} \qquad a=1,\ b=-37,\ c=201$$

$\boxed{\text{Eigen values} \quad \lambda_1 = 30.38, \quad \lambda_2 = 6.62}$

Arrange in descending orders $\lambda_1 > \lambda_2 > \dots \lambda_n$

Hence $\lambda_1 = 30.38$

$\lambda_2 = 6.62$

NOW, compute Eigen vectors using the above Eigen values λ_1 & λ_2.

$\Rightarrow$ Eigen vector $[V_1 \quad V_2]$

Let $V_1 = \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix}$ $\quad V_2 = \begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix}$

It will be computed as:

$$[S - \lambda_1 I] \cdot V_1 = 0$$

$$\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \lambda_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_1 \end{bmatrix} \cdot \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} (14-\lambda_1) & (-11) \\ (-11) & (23-\lambda_1) \end{bmatrix} \cdot \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 0$$

$$\Rightarrow (14-\lambda_1) V_{11} - 11 V_{12} = 0 \quad — \text{①}$$

$$-11 V_{11} + (23-\lambda_1) V_{12} = 0 \quad — \text{②}$$

from ① $\Rightarrow \dfrac{V_{11}}{11} = \dfrac{V_{12}}{(14-\lambda_1)} = A$ (Assume)

Let $A = 1$

$$\frac{V_{11}}{11} = \frac{V_{12}}{(14-\lambda_1)} = 1$$

$$\Rightarrow \frac{v_{11}}{11} = 1 \qquad \frac{v_{12}}{14-\lambda_1} = 1$$

$$\Rightarrow \boxed{v_{11} = 11} \qquad \Rightarrow v_{12} = 1(14-\lambda_1)$$

$$= 14 - 30.38$$

$$\boxed{v_{12} = -16.38}$$

$\therefore$ Eigen vector for λ_1 is $v_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}$

$$v_1 = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix}$$

Now normalize this v_1

$$v_1 = \begin{bmatrix} 11/\sqrt{11^2+(-16.38)^2} \\ -16.38/\sqrt{11^2+(-16.38)^2} \end{bmatrix}$$

$$\boxed{v_1 = \begin{bmatrix} 0.5575 \\ -0.8302 \end{bmatrix}} \quad \text{for} \quad \boxed{\lambda_1 = 30.38}$$

Now compute Eigen vector v_2 for $\lambda_2 = 6.62$

$$[S - \lambda_2 I]\begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

$$\begin{bmatrix} (14-\lambda_2) & (-11) \\ (-11) & (23-\lambda_2) \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

$(14-\lambda_2) v_{21} - 11 v_{22} = 0$ — 3

$(-11) v_{21} + (23-\lambda_2) v_{22} = 0$ — 4

from 3

$$(14-\lambda_2) v_{21} - 11 v_{22} = 0$$

$$\Rightarrow (14-\lambda_2) v_{21} = 11 v_{22}$$

$$\Rightarrow \frac{v_{21}}{11} = \frac{v_{22}}{(14-\lambda_2)} = B \quad \text{(Assume)}$$

Let $B = 1$

$$\Rightarrow \frac{v_{21}}{11} = 1, \quad \frac{v_{22}}{(14-\lambda_2)} = 1$$

$$\Rightarrow \boxed{v_{21} = 11} \quad v_{22} = 1(14-\lambda_2)$$

$$= 14 - 6.62$$

$$\boxed{v_{22} = 7.38}$$

Eigen vector for λ_2 is $v_2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 11 \\ 7.38 \end{bmatrix}$

Now normalize this $\underline{v_2}$

$$v_2 = \begin{bmatrix} 11 / \sqrt{11^2 + 7.38^2} \\ 7.38 / \sqrt{11^2 + 7.38^2} \end{bmatrix}$$

$$\boxed{v_2 = \begin{bmatrix} 0.8308 \\ 0.5574 \end{bmatrix}} \quad \text{for} \quad \boxed{\lambda_2 = 6.62}$$

Step 5 :- Compute New dataset.

feature	sample1	sample2	sample3	sample4
a	4	8	13	7
b	11	4	5	14

⇒

1^{st} PC	P_{11}	P_{12}	P_{13}	P_{14}

$$P_{11} = V_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5575 & -0.8302 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$\boxed{P_{11} = -4.305}$$

$$P_{12} = V_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 & -0.8302 \end{bmatrix} \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$$

$$\boxed{P_{12} = 3.735}$$

$$P_{13} = V_1^T \begin{bmatrix} 13-8 \\ 5-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 & -0.8302 \end{bmatrix} \begin{bmatrix} 5 \\ -3.5 \end{bmatrix}$$

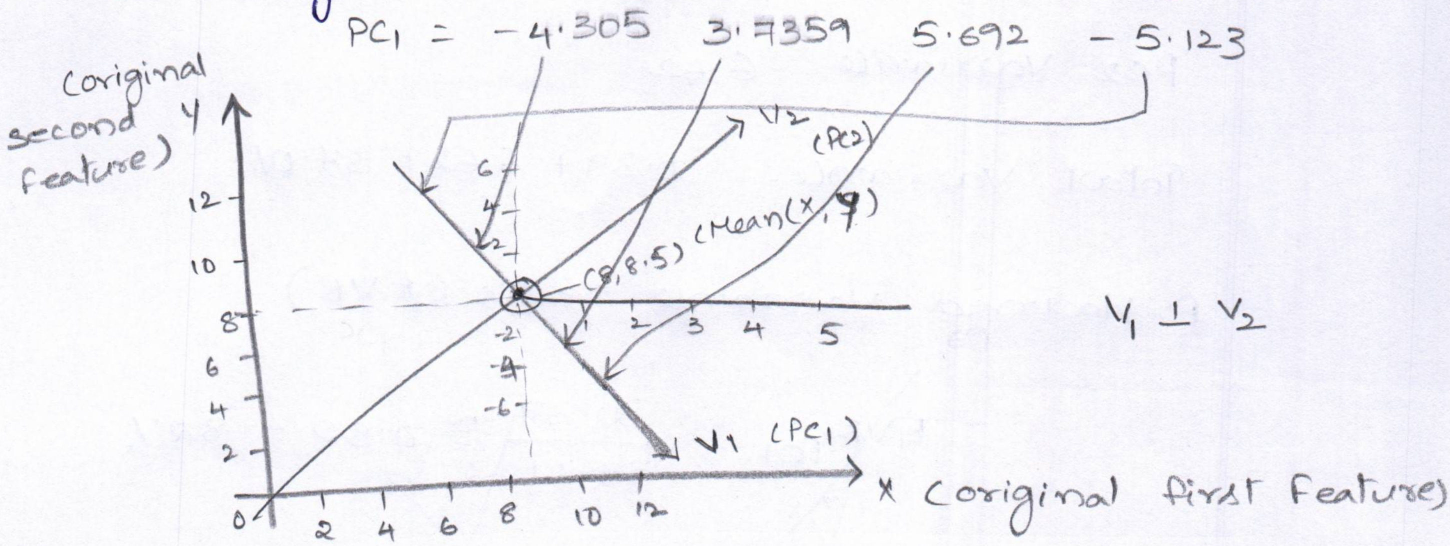
$$\boxed{P_{13} = 5.692}$$

$$P_{14} = V_1^T \begin{bmatrix} 7-8 \\ 14-8.5 \end{bmatrix} = \begin{bmatrix} 0.5575 & -0.8302 \end{bmatrix} \begin{bmatrix} -1 \\ 5.5 \end{bmatrix}$$

$$\boxed{P_{14} = -5.123}$$

$$\boxed{PC_1 = -4.305, \ 3.735, \ 5.692, \ -5.123}$$

similarlly we can compute PC_2 using V_2



How many PCA's are need for any data?

Eigen values are used to find out which PCA has a maximum variance.

PC_2 using V_2

$$\boxed{PC_2 = -1.9297, \ -2.5083, \ 2.2031, \ 2.2349}$$

variance of PCA's

$$Var(PC_1) = \frac{1}{4-1}\left[(-4.305)^2 + (3.7359)^2 + (5.692)^2 + (-5.123)^2\right]$$

$$= \frac{1}{3}\cdot[91.16]$$

$$\boxed{Var(PC_1) \approx 30.39}$$

$$var(PC_2) = \frac{1}{4-1}\left[(-1.9297)^2 + (-2.5083)^2 + (2.2031)^2 + (2.2349)^2\right]$$

$$\boxed{Var(PC_2) \cong 6.621}$$

PC1 variance = 30.39

PC2 variance = 6.62

Total variance = 30.39 + 6.62 = 37.01

Explained variance ratio (EVR)

$$EVR_{PC1} = \frac{30.39}{37.01} \simeq 0.82 = 82\%$$

$$EVR_{PC2} = \frac{6.62}{37.01} \simeq 0.18 = (18\%)$$

⇒ PC_1 explains 82% of the data's information.

PC2 " 18% " " " "

How many PCA's required for this is data

is PC_1 & PC_2

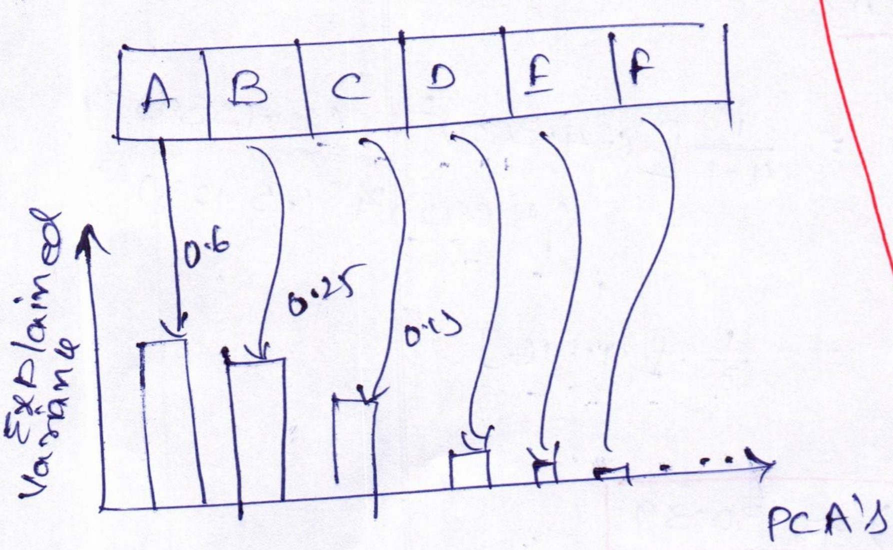
Case A: - Dimensionality reduction

PC1 alone captures 82% data

reduces 2D - 1D without much loss

Case B: If we need to retain 100% keep two PC's

Final Answer: Minimum PC's



~~in our case 82+18 ⇒ 100%.~~ 2 PC's are enough.