principal Component Analysis (PCA)

principal Component Analysis, or PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Step1: Standardization

The aim of this step is to standardize the range of continuous initial variables so that each one of them contributes equally to the analysis

we compute mean of feature variables x & y

Step2: Covariance matrix Computation

The Covariance matrix is a n×n symmetric matrix (where n is the number of dimensions) that has as entries the covariance associated with all possible pairs of the initial variables

for example for a 2-dimensional data set with 2 variables x, y. the covariance matrix is a 2×2 matrix of this form

$$\begin{array}{c|cc} & x & y \\ \hline x & \text{Cov}(x,x) & \text{Cov}(x,y) \\ y & \text{Cov}(y,x) & \text{Cov}(y,y) \end{array}$$

Step 3: Compute the Eigen values and Eigen vectors of Covariance matrix to identify the principal Components

principal Components are constructed in such a manner that the first principal Component accounts for the largest possible variance in the data set.

By Ranking our Eigen vectors in order of their eigen values, highest to lowest, we get the principal Components in order of Significance.

Ex: let us suppose that our data set is a 2-dimensional with 2 variables x, y and that the eigen vectors and eigen values of the Covariance matrix are as follows

$$V_1 = \begin{pmatrix} 0.6778 \\ 0.7351 \end{pmatrix} \qquad \lambda_1 = 1.2840$$

$$V_2 = \begin{pmatrix} -0.7351 \\ 0.6778 \end{pmatrix} \qquad \lambda_2 = 0.049$$

If we rank the eigen values in descending order, we get $\lambda_1 > \lambda_2$. which means that the eigen vector that corresponds to the first principal component (PC 1) is V_1 and the one that corresponds to the Second Component (PC2) is V_2

$$\therefore \quad V_1 = \frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{1.2840}{(1.2840+0.049)} = 96\% \text{ (information)}$$

$$V_2 = \frac{\lambda_2}{\lambda_1+\lambda_2} = \frac{(0.049)}{(1.2840+0.049)} = 4\% \text{ of variance (information) of the data}$$

The feature vector is simply a matrix that has as Columns the eigen vectors of the Components that we decide to keep.

Final data Set = Feature vector^T x (Standardized original data Set)^T

Given the following data use PCA (principal Component Analysis) to reduce the dimension from 2 to 1.

Feature	Example 1	Example 2	Example 3	Example 4
x	4	8	13	7
y	11	4	5	14

Step_1 Data Set

Feature	ex.1	Ex.2	Ex3	Ex4
x	4	8	13	7
y	11	4	5	14

Number of feature $n = 2$

Number of samples $N = 4$

Step 2: Computation of Mean of variables

$$\bar{x} = \frac{4+8+13+7}{4} = \frac{32}{4} = 8$$

$$\bar{y} = \frac{11+4+5+14}{4} = \frac{34}{4} = 8.5$$

Step 3: Computation of Covariance matrix

we need to write ordered pairs

$$\begin{array}{c|cc} & x & y \\ \hline x & & \\ y & & \end{array}$$

(x, x) (x, y), (y, x), (y, y)

Covariance of all the ordered pairs

$$\text{Cov}(x, x) = \frac{1}{(N-1)} \sum_{K=1}^{N} (x_{iK} - \bar{x}_i)(x_{jk} - x_j)$$

$$\boxed{\text{Cov}(x,x) = \frac{1}{(N-1)} \sum_{K=1}^{N} (x_i - \bar{x})^2} = \frac{1}{(4-1)} \left[(4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2 \right]$$

$$= \frac{1}{3}(16 + 0 + 25 + 1) = \frac{1}{3}(42)$$

$$= 14$$

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

x	4	8	13	7
y	11	4	5	14

$$= \frac{1}{4}\left[-4(3.5)+0+5(-3.5)+(-1)(5.5)\right]$$

$$= \frac{1}{3}\left[-33\right] = -11$$

$$cov(y,x) = cov(x,y) = -11$$

$$cov(y,y) = \frac{1}{(4-1)}\sum_{K=1}^{N}(y_i-\bar{y})^2$$

$$= \frac{1}{3}\left[(11-8.5)^2+(4-8.5)^2+(5-8.5)^2+(14-8.5)^2\right]$$

$$= \frac{1}{3}\left[(2.5)^2+(-4.5)^2+(-3.5)^2+(5.5)^2\right]$$

$$= 23$$

Covariance matrix $n\times n\ (=2\times 2)$

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

Step 4: Eigen value $\Rightarrow$ Eigen vectn $\Rightarrow$ Nomalized eigen vetor

$\therefore$ Characteristic Equation is $|A - I\lambda| = 0$

$$\begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = (14-\lambda)(23-\lambda) - 121 = 0$$

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

$$\lambda_1 = 30.3849, \quad \lambda_2 = 6.6151$$

$$\lambda_1 > \lambda_2$$

Eigen vectn of $\lambda_1 = 30.3849$

$$(S-\lambda_1 I)U_1 = 0 \quad \begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} (14-\lambda_1) & -11 \\ -11 & 23-\lambda_1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(14-\lambda_1)\, u_1 - 11 u_2 = 0$$

$$-11 u_1 + (23-\lambda_1)\, u_2 = 0$$

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda_1} = t \Rightarrow \quad t=1, \quad u_1 = 11, \quad u_2 = 14-\lambda_1$$

Eigen vector u_1 of λ_1 $= \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix} = \begin{bmatrix} 11 \\ 14-30.3849 \end{bmatrix} = \begin{bmatrix} 11 \\ -16.3849 \end{bmatrix}$

Normalize the Eigen Vector

$$e_1 = \frac{u_1}{\|u_1\|} = \frac{1}{\sqrt{(11)^2 + (-16.3849)^2}} \begin{pmatrix} 11 \\ -16.3849 \end{pmatrix} = \begin{bmatrix} 0.5579 \\ -0.8303 \end{bmatrix}$$

unit Eigen vector

Illy we can find second normalized vector as $e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$

Step5: Derive new data set

	Ex1	Ex2	Ex3	Ex4
first principal Component PC	P_{11}	P_{12}	P_{13}	P_{14}

$$P_{11} = e_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5574, -0.8303 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix} = -4.3052$$

$$P_{12} = e_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} = \begin{bmatrix} 0.5574, -0.8303 \end{bmatrix} \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$$

$$= \begin{bmatrix} 3.7361 \end{bmatrix}$$

$$P_{13} = e_1^T \begin{bmatrix} 13-8 \\ 5-8.5 \end{bmatrix} = \begin{bmatrix} 0.5574, -0.8303 \end{bmatrix} \begin{bmatrix} 5 \\ -3.5 \end{bmatrix}$$

$$= 5.6928$$

$$P_{14} = e_1^T \begin{bmatrix} 7-8 \\ 14-8.5 \end{bmatrix} = \begin{bmatrix} 0.5574, -0.8303 \end{bmatrix} \begin{bmatrix} -1 \\ 5.5 \end{bmatrix}$$

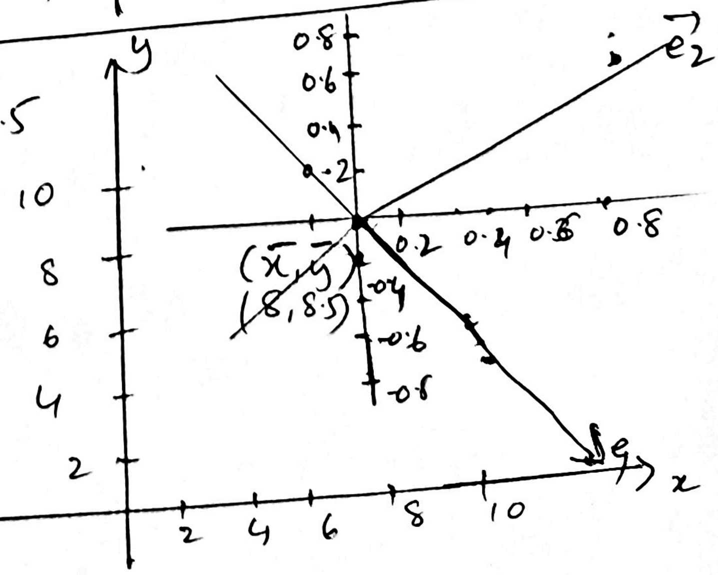
$$= -5.1238$$

	Ex1	Ex2	Ex3	Ex.4
PC	-4.3052	3.7361	5.6928	-5.1238

$\bar{x}_1 = 8 \quad \bar{y} = 8.5$

$$e_1 = \begin{pmatrix} 0.55 \\ -0.83 \end{pmatrix}$$

$$e_2 = \begin{pmatrix} 0.83 \\ 0.55 \end{pmatrix}$$



$e_1 \perp e_2$

① Given the following data use (principal component Analysis) to reduce the dimension from 2 to 1.

Step 1:

Feature	Example 1	Example 2	Example 3	Example 4
x	2	1	0	-1
y	4	3	1	0.5

Step 2 $\bar{x} = \frac{2}{4} = 0.5$ $\bar{y} = \frac{8.5}{4} = 2.125$

Table:

x	y	A $(x_i - \bar{x})$	B $(y_i - \bar{y})$	A.B	A^2	B^2
2	4	1.5	1.875	2.8125	2.25	3.5156
1	3	0.5	0.875	0.1375	0.25	0.07562
0	1	-0.5	-2.125	0.5625	0.25	1.2656
-1	0.5	-1.5	-1.625	2.4375	2.25	2.6406
2	8.5			6.25	5	7.4974

Step 3: $\text{cov}(x, x) \Rightarrow \boxed{\sum_{i=1}^{n} \frac{(x - x_i)(y_i - y_i)}{n-1} = \text{cov}(x, y)}$

$$\text{cov}(x, x) = \sum_{i=1}^{n} \frac{(x - \bar{x}_i)^2}{(n-1)} = \sum_{i=1}^{n} \frac{A^2}{3} = \frac{5}{3} = 1.67 \quad \text{—①}$$

$$\text{cov}(y, y) = \sum_{i=1}^{n} \frac{(y - \bar{y}_i)^2}{(n-1)} = \sum_{i=1}^{n} \frac{B^2}{3} = \frac{7.4974}{3} = 2.499 \quad \text{—②}$$

$$\text{cov}(x, y) = \sum_{i=1}^{n} \frac{(x - \bar{x}_i)(y - y_i)}{(n-1)} = \sum_{i=1}^{n} \frac{AB}{3} = \frac{6.25}{3} = 2.08$$

$\because \text{cov}(y, x) = 2.083$

$$S = \text{Covariance matrix } \begin{array}{c|cc} & x & y \\ \hline x & (x,x) & (x,y) \\ y & (y,x) & (y,y) \end{array}$$

$$= \begin{bmatrix} 1.67 & 2.083 \\ 2.083 & 2.49 \end{bmatrix}$$

$$\simeq S = \begin{bmatrix} 1.67 & 2.08 \\ 2.08 & 2.49 \end{bmatrix}$$

Characteristic Equation

Step 4 $|S - I\lambda| = \begin{vmatrix} 1.67-\lambda & 2.08 \\ 2.08 & 2.49-\lambda \end{vmatrix} = 0$

$$(1.67-\lambda)(2.49-\lambda)-(2.08)^2 = 0$$

$$\lambda^2 - 4.16\lambda - 0.1681 = 0$$

$$(\lambda + 0.04)(\lambda - (4.20)) = 0$$

The Eigen values of the matrix A are given by $\lambda = -0.04$, $\lambda_2 = 4.20$

The Eigen vector for $\lambda = 4.20$ is

$(A - I\lambda)X = 0$ & $(S - I\lambda)X = 0$

$$\begin{pmatrix} 1.67-4.20 & 2.08 \\ 2.08 & 2.49-4.20 \end{pmatrix} X = \begin{pmatrix} -2.53 & 2.08 \\ 2.08 & -1.71 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$R_2 \to R_2(2.53) + (2.08)R_1$

Solving this we get $V_1 = \begin{pmatrix} 0.8221 \\ 1 \end{pmatrix}$

$\therefore$ the Eigen vector for $\lambda = -0.04$ is

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

$$\begin{pmatrix} 1.71 & 2.08 \\ 2.08 & 2.53 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad R_2 \to 1.71R_2 - 2.08R_1$$

on solving this we get

Eigen vector $V_2 = \begin{pmatrix} -1.2163 \\ 1 \end{pmatrix}$

orthonormal vector

$$e_1 = \frac{v_1}{\|v_1\|} = \frac{1}{\sqrt{1+(0.8221)^2}} \begin{pmatrix} 0.8221 \\ 1 \end{pmatrix}$$

$$= \frac{1}{\sqrt{1.6758}} \begin{pmatrix} 0.8221 \\ 1 \end{pmatrix} = 0.7724 \begin{pmatrix} 0.8221 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.6350 \\ 0.7724 \end{pmatrix}$$

$$e_2 = \frac{v_2}{\|v_2\|} = \frac{1}{\sqrt{1^2+(-1.2163)^2}} \begin{pmatrix} -1.2163 \\ 1 \end{pmatrix}$$

$$= \frac{1}{\sqrt{2.4793}} \begin{pmatrix} -1.2163 \\ 1 \end{pmatrix} = 1.5745 \begin{pmatrix} -1.2163 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} -1.9151 \\ 1.5745 \end{pmatrix}$$

Step 5: Derive the new data set

	Ex 1	Ex 2	Ex 3	Ex 4
First principal Component PC1	P_{11}	P_{12}	P_{13}	P_{14}

$$P_{11} = e_1^T \begin{pmatrix} x_1-\bar{x} \\ y_1-\bar{y} \end{pmatrix} = \begin{pmatrix} 0.6350 & 0.7724 \end{pmatrix} \begin{pmatrix} 1.5 \\ 1.85 \end{pmatrix} = 2.3814$$

$$P_{12} = e_1^T \begin{pmatrix} x_2-\bar{x} \\ y_2-\bar{y} \end{pmatrix} = \begin{pmatrix} 0.6350 & 0.7724 \end{pmatrix} \begin{pmatrix} 0.5 \\ 0.275 \end{pmatrix} = 1.7464$$

$$P_{13} = e_1^T \begin{pmatrix} x_3-\bar{x} \\ y_3-\bar{y} \end{pmatrix} = \begin{pmatrix} 0.6350 & 0.7724 \end{pmatrix} \begin{pmatrix} -0.5 \\ -1.125 \end{pmatrix} = -1.1767$$

$$P_{14} = e_1^T \begin{pmatrix} x_4-\bar{x} \\ y_4-\bar{y} \end{pmatrix} = \begin{pmatrix} 0.6350 & 0.7724 \end{pmatrix} \begin{pmatrix} -1.5 \\ -1.625 \end{pmatrix} = -2.20765$$

	Ex 1	Ex 2	Ex 3	Ex 4
PG :	2.3816	1.7464	-1.1767	-2.20765

$$e_1 = \begin{pmatrix} 0.63 \\ 0.77 \end{pmatrix}$$

$$e_2 = \begin{pmatrix} -1.91 \\ 1.57 \end{pmatrix}$$

$(\bar{x}, \bar{y}) = (0.5, 2.215)$