Experiment No. 6 Date: 13/03/26

Title:- Principle Component analysis On 2D data.

Aim: to Implement Principal Component analysis (PCA) on a 2D dataset, reduce its dimensionality while preserving significant variance and visualize the transformed data.

Theory:

Principal Component Analysis:

PCA is a statistical technique used for dimensionality reduction while retaining as much variance as possible in the data. It transforms the data into new coordinate system where the most significant features (principal component) capture the highest variance.

PCA is widely used in machine learning, Pattern recognition, Image processing and data visualization to reduce complexity, improve computational efficiency and remove redundant information.

For the dataset having d dimensions

for the Mean centered the data.

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$X_{centered} = X - \bar{X}$$

where x_{ij} represented the value of the j^{th} feature for the i^{th} Sample.

for the relationship b/w different features

$$C = \frac{1}{n-1} X^T_{centered} X_{centered}$$

where c is the d×d symmetric matrix where c_{ij} represents the covariance b/w features i and feature j.

Algorithms

→ Load the dataset.

Given a dataset X of size n×d, where each row represent a sample, and each column represent feature.

⇒ Mean Centering the data.

Compute the mean of the each features

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

Subtract the mean from each data point to obtain the mean centered data

$$X_{centered} = X - \bar{X}$$

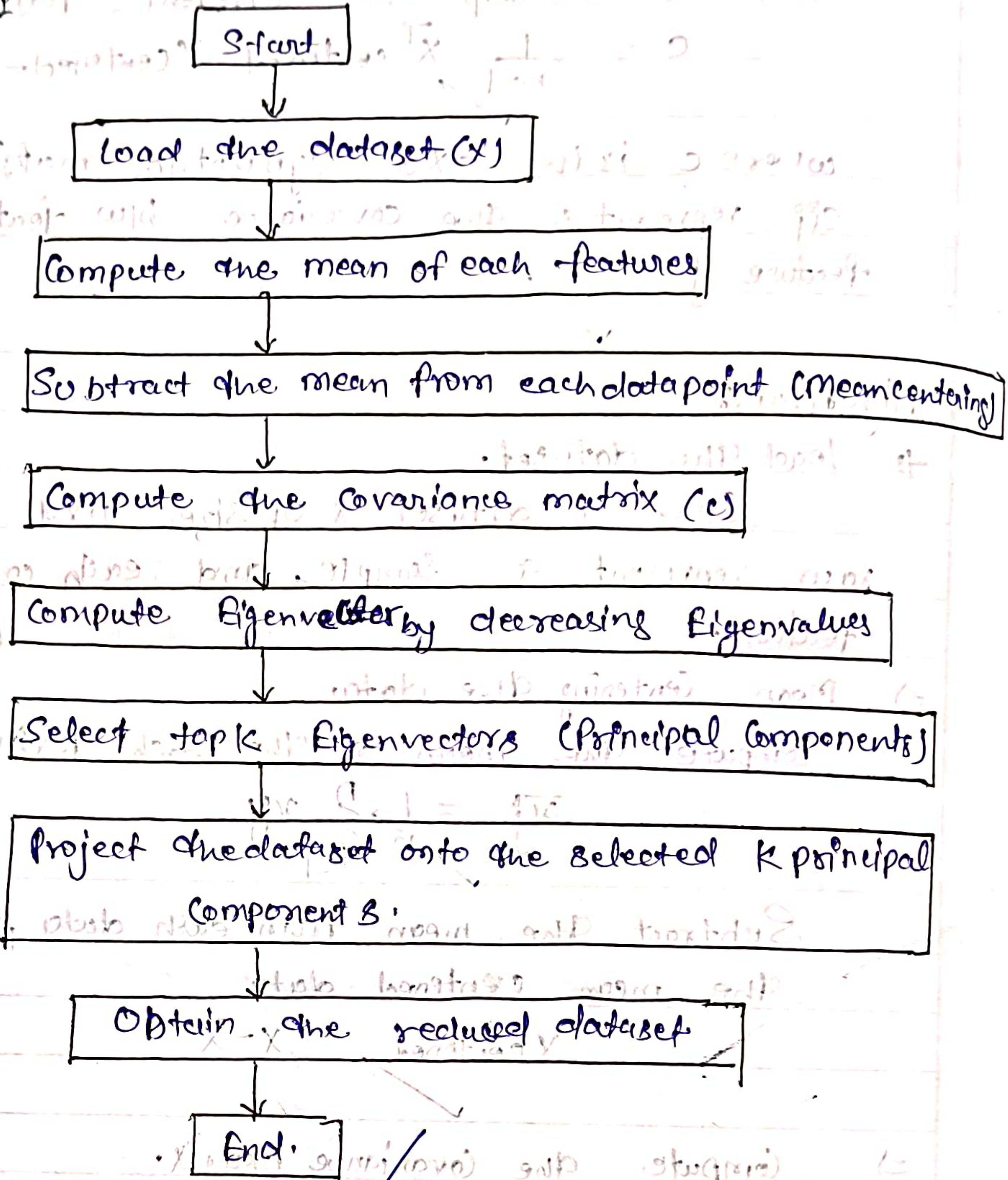
⇒ Compute the Covariance Matrix.

Calculate the Covariance matrix C.

$$C = \frac{1}{n-1} X^T_{centered} X_{centered}$$

The covariance matrix is of size d×d.

flowchart:


Start.
Load the dataset (X)
Compute the mean of each features
Subtract the mean from each datapoint (Mean centering)
Compute the covariance matrix (C)
Compute Eigenvector by decreasing Eigenvalues
Select top k Eigenvectors (Principal Components)
Project the dataset onto the selected k principal components.
Obtain the reduced dataset
End.

⇒ Compute Eigenvalues and Eigen vectors:
Solve the eigen values equations.

$$CV = \lambda V$$

Obtain the eigenvalues λ and corresponding eigenvalues V.

⇒ Sort Eigenvalues Eigenvectors by Eigenvalues.
→ Arrange eigen~~value~~ vectors in descending order of their eigenvalues.
Select the top k eigen~~values~~ vectors corresponding to the top highest k eigen values to form the projection matrix $P_K = [v_1, v_2, v_3 \cdots - v_k]$

⇒ Project Data Onto Principle Component
Transform the dataset into the new space using

$$X_{reduced} = X_{centered} P_K.$$

if $K=d$, all components are retained.
if $K<d$, dimensionality is reduced.

⇒ Output the transformed Dataset
The reduced dataset X reduced is now of size $n \times k$ where k is the no of principal component retained.

Results

$$mean = [1.81, 1.91]$$

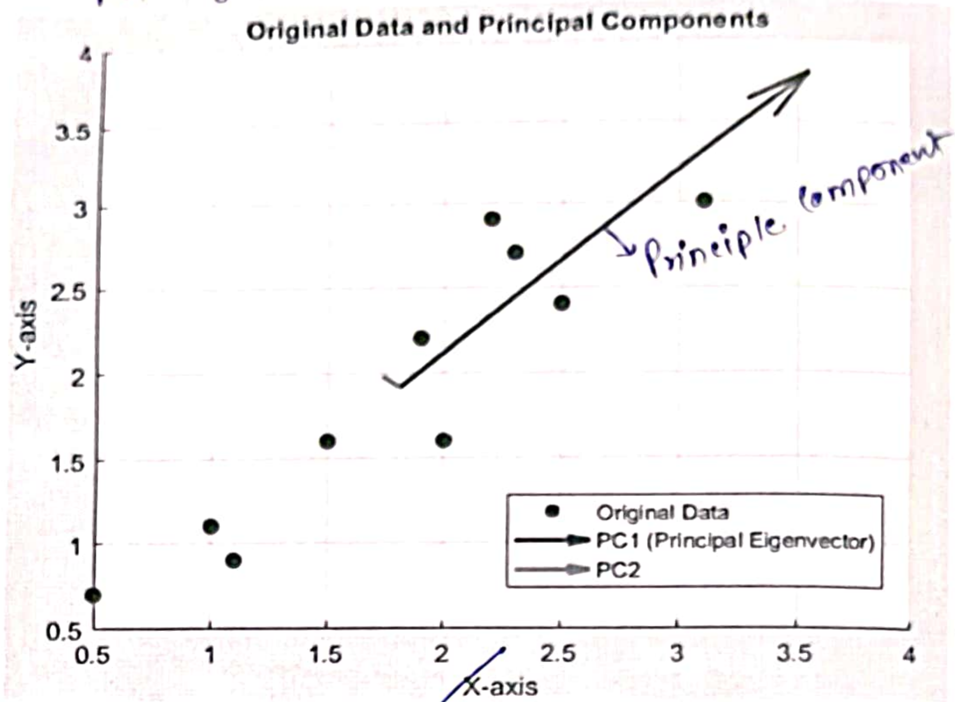
$$\text{Covariance Matrix} = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

$$\text{Eigenvalues} = [1.2840, 0.0491]$$

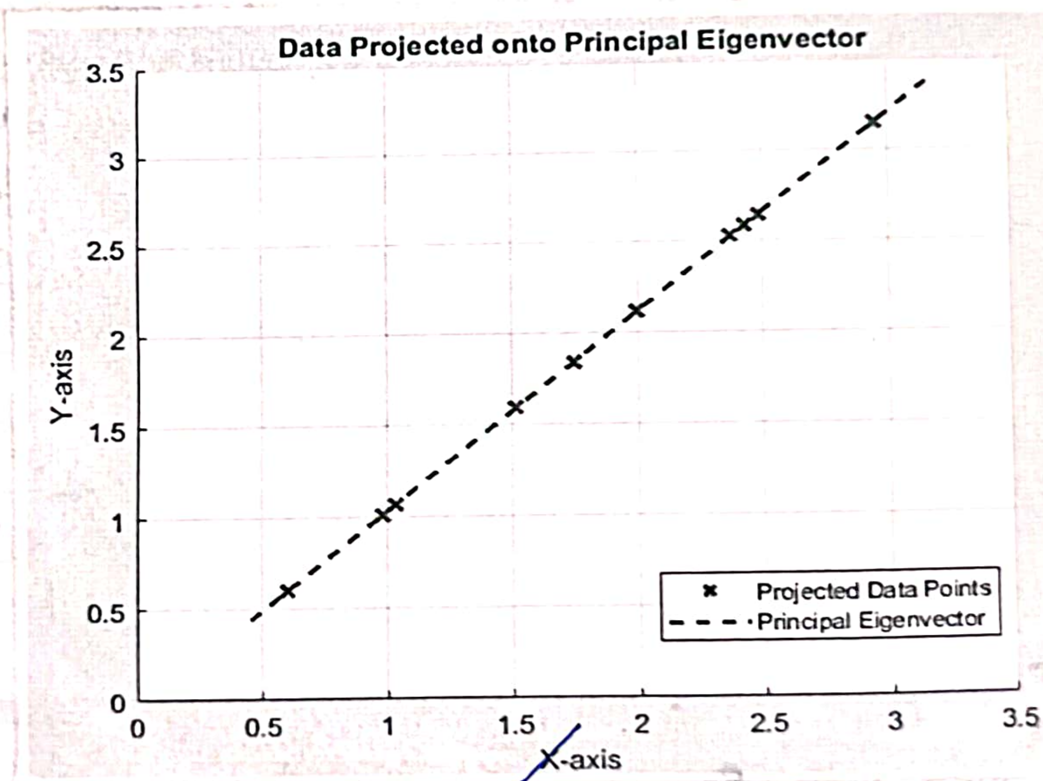
$$\text{Eigenvectors} = [0.6779 \quad -0.7352]^T [0.7352, 0.6779]$$

Result:

taking 10 samples.



After dimention reduction 2D to 1D



	2D data	1D data
	(1.5, 1.6)	-0.1380
Data:	(2.5, 2.7)	0.8828
	(0.5, 0.7)	-1.776
	(2.2, 2.9)	0.9922
	(1.9, 2.2)	0.2712
	(3.1, 3.0)	1.6758
	(2.3, 2.7)	0.9129
	(2.0, 1.6)	-0.0991
	(1.0, 1.1)	-1.1976
Discussion:	(1.1, 0.9)	-1.2238

⇒ The points are randomly distributed but may have a correlation b/w X and y.

⇒ Mean centering: The data is shifted so that its mean is at the origin.

⇒ The covariance matrix reveals the relationship b/w the two variables.

⇒ PCA finds two orthogonal direction (Principal components) we can project component to reduce it to 1D

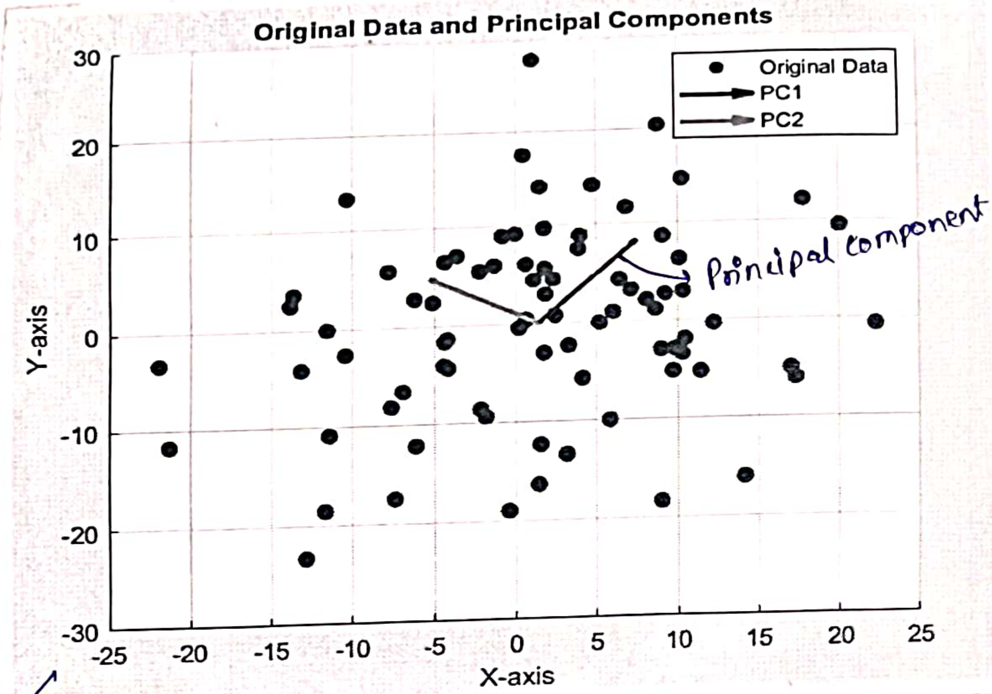
⇒ The first principal component captures the highest variance direction.

⇒ The second principal component is perpendicular to the first and captures less variance.

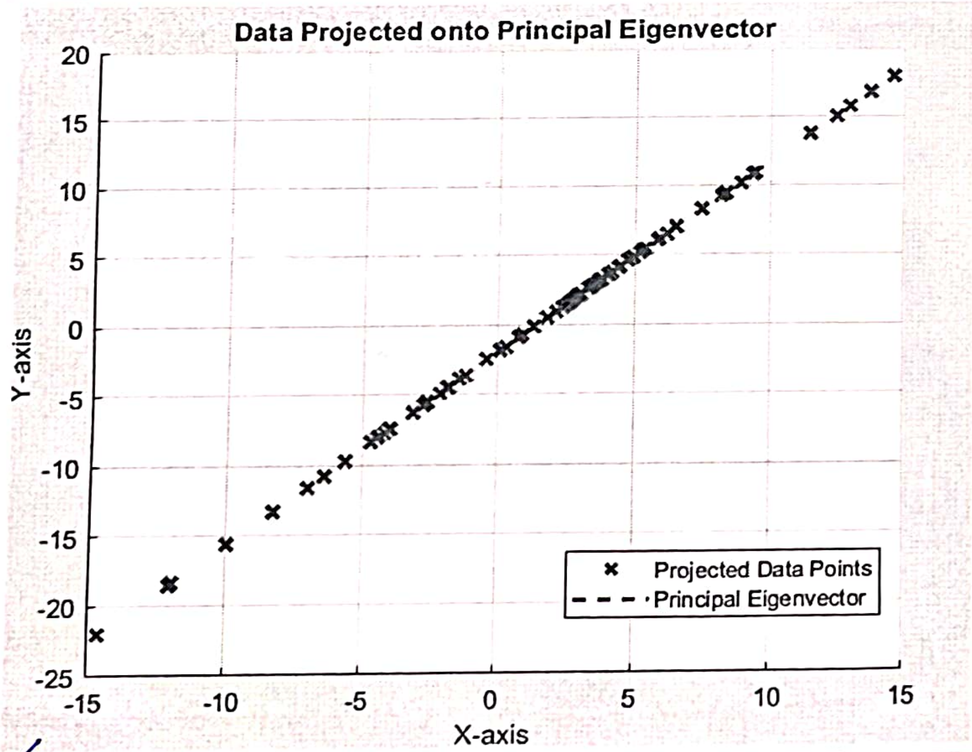
⇒ If x and y are highly correlated, PCA aligns along the diagonal; otherwise it distributes orthogonal.

⇒ The variance along the second component is low, making it redundant for most applications.

Result→ taking the random - 80 datasample in 2D plane



after dimention reduction 2D to 1D



Conculsion:

⇒ PCA is effective for dimensionality reduction, especially when the data has correlated features.

⇒ It helps in feature extraction and noise reduction by removing less significant component.

⇒ The first principal component is the most important and can be used to approximate the data in a lower-dimention space.

⇒ PCA is useful for visualization, enables data representation in a reduced dimension without much information loss.

⇒ If the dataset is Uniform variance in all directions, PCA does not provide significant reduction benefits.

⇒ The eigenvalues indicates how much variance each component retains, guiding decision on the no of components to keep.

⇒ PCA is widely used in data compression, pattern recognition and noise fitting.

⇒ PCA should be used carefully when data interpretation is necessary, as transformed features may lose their original meaning.

6-08-25