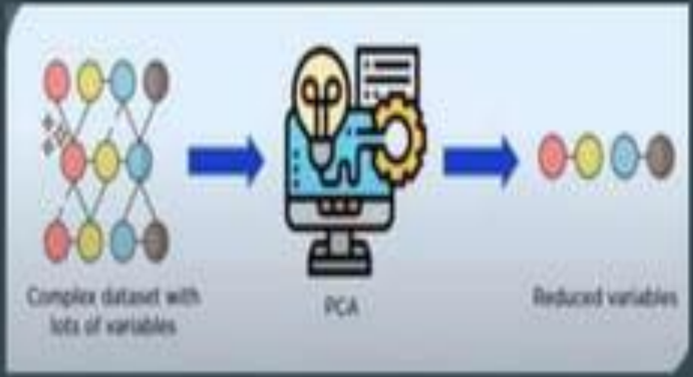
Principal Component Analysis (PCA)

...

Introduction

Principal Component Analysis, commonly referred to as PCA, is a powerful mathematical technique used in data analysis and statistics. At its core, PCA is designed to simplify complex datasets by transforming them into a more manageable form while retaining the most critical information.



- reducing the dimensionality of dataset
- Increasing interpretability without losing information

PCA (Principal Component Analysis)

PCA (Principal Component Analysis) is a **dimensionality reduction** technique used in data analysis and machine learning.

It helps you to reduce the number of features in a dataset while keeping the most important information.

It changes your original features into new features these new features don't overlap with each other and the first few keep most of the important differences found in the original data.

Dimensionality Reduction

Dimensionality reduction refers to the techniques that reduce the number of input variables in a dataset.

Why DR?

- Less dimensions for a given dataset means less computation or training time
- Redundancy is removed after removing similar entries from the dataset
- Data Compression (Reduce storage space)
- It helps to find out the most significant features and skip the rest
- Leads to better human interpretations

	Student 1	Student 2	Student 3	Student 4	...
Math	95	88	93	75	...
Reading	96	79	98	81	...

For example, the samples could be students in high school and the variables could be test scores in math and reading…

Why PCA?

- Dimensionality Reduction
- Noise Reduction
- Visualization
- Feature Engineering
- Overfitting Problem
- Data Compression
- Machine Learning Processing

How Principal Component Analysis Works

PCA uses **linear algebra** to transform data into new features called principal components.

It finds these by calculating **eigenvectors (directions)** and **eigenvalues (importance)** from the **covariance matrix.**

PCA selects the top components with the **highest eigenvalues** and projects the data onto them simplify the dataset.

Important Terminologies

- Variance
- Covariance
- Eigenvalues
- Eigenvectors
- Principle Component

Important Terminologies (Variance)

- **Variance** is the sum of squares of differences between all numbers and means.

- Variance (σ^2) = (Sum of the squared differences from the mean) / (Total number of values)
- In mathematical notation: $\sigma^2 = \Sigma(x - \mu)^2 / (n)$

Here:

- μ is the mean of independent features

-Mean (μ) = (Sum of all values) / (Total number of values)

Important Terminologies (Covariance)

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

Important Terminologies (Covariance)

The formula for the covariance (Cov) between two random variables X and Y, each with N data points, is as follows:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

Where:

- Cov(X,Y) is the covariance between X and Y.
- N is the number of data points.
- Xi and Yi represent individual data points for X and Y, respectively.

Principal Component Analysis – Algorithm

Step 1. Data

- We consider a dataset having n features or variables denoted by $X_1; X_2; \ldots ; X_n$.
- Let there be N examples.
- Let the values of the i^{th} feature X_i be $X_{i1}; X_{i2}; \ldots ; X_{iN}$

Features	Example 1	Example 2	$\cdots$	Example N
X_1	X_{11}	X_{12}	$\cdots$	X_{1N}
X_2	X_{21}	X_{22}	$\cdots$	X_{2N}
$\vdots$				
X_i	X_{i1}	X_{i2}	$\cdots$	X_{iN}
$\vdots$				
X_n	X_{n1}	X_{n2}	$\cdots$	X_{nN}

Principal Component Analysis – Algorithm

Step 2. Compute the means of the variables

Features	Example 1	Example 2	$\cdots$	Example N
X_1	X_{11}	X_{12}	$\cdots$	X_{1N}
X_2	X_{21}	X_{22}	$\cdots$	X_{2N}
$\vdots$				
X_i	X_{i1}	X_{i2}	$\cdots$	X_{iN}
$\vdots$				
X_n	X_{n1}	X_{n2}	$\cdots$	X_{nN}

$$\bar{X}_i = \frac{1}{N}(X_{i1} + X_{i2} + \cdots + X_{iN})$$

Principal Component Analysis – Algorithm

Step 3. Calculate the covariance matrix

Features	Example 1	Example 2	$\cdots$	Example N
X_1	X_{11}	X_{12}	$\cdots$	X_{1N}
X_2	X_{21}	X_{22}	$\cdots$	X_{2N}
$\vdots$				
X_i	X_{i1}	X_{i2}	$\cdots$	X_{iN}
$\vdots$				
X_n	X_{n1}	X_{n2}	$\cdots$	X_{nN}

$$\mathrm{Cov}(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)$$

Covariance matrix

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & & & \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

Principal Component Analysis – Algorithm

Step 4. Calculate the eigenvalues and eigenvectors of the covariance matrix

i. **Set up the equation:** This is a polynomial equation of degree **n** in. It has **n** real roots and these roots are the eigenvalues of **S**

$\lambda_1 \; \lambda_2 \ldots \lambda_n$

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

ii. **If $\lambda = \lambda'$ is an eigenvalue**, then the corresponding eigenvector is a vector

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{such that} \quad (S - \lambda' I)U = 0$$

Principal Component Analysis – Algorithm

Step 4. Calculate the eigenvalues and eigenvectors of the covariance matrix

iii. **We now normalize the eigenvectors**. Given any vector X we normalize it by dividing X by its length. The length (or, the norm) of the vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is defined as

$$\|X\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

We compute the n normalised eigenvectors $e_1, e_2, \ldots, e_n$ by

$$e_i = \frac{1}{\|U_i\|} U_i, \quad i = 1, 2, \ldots, n.$$

Principal Component Analysis – Algorithm

Step 5. Derive new data set

- Order the eigenvalues from highest to lowest.
- The unit eigenvector corresponding to the largest eigenvalue is the first principal component.

i) Let the eigenvalues in descending order be $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and let the corresponding unit eigenvectors be $e_1, e_2, \ldots, e_n$.

ii) Choose a positive integer p such that $1 \leq p \leq n$.

iii) Choose the eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and form the following $p \times n$ matrix (we write the eigenvectors as row vectors):

Principal Component Analysis – Algorithm

Step 5. Derive new data set

- Order the eigenvalues from highest to lowest.
- The unit eigenvector corresponding to the largest eigenvalue is the first principal component.

$n = 10 \quad p = 5$

i) Let the eigenvalues in descending order be $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and let the corresponding unit eigenvectors be $e_1, e_2, \ldots, e_n$.

ii) Choose a positive integer p such that $1 \leq p \leq n$.

iii) Choose the eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and form the following $p \times n$ matrix (we write the eigenvectors as row vectors):

$$F = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_p^T \end{bmatrix}$$

Principal Component Analysis – Algorithm

Step 5. Derive new data set

iv) We form the following $n \times N$ matrix:

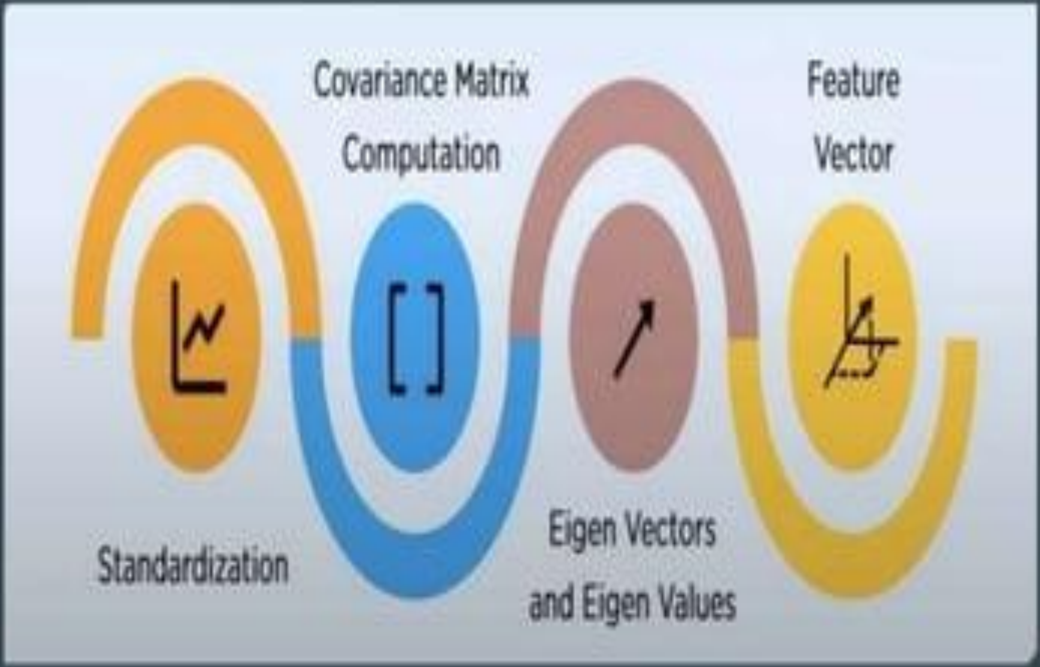
$$X = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_1 & \cdots & X_{1N} - \bar{X}_1 \\ X_{21} - \bar{X}_2 & X_{22} - \bar{X}_2 & \cdots & X_{2N} - \bar{X}_2 \\ \vdots & & & \\ X_{n1} - \bar{X}_n & X_{n2} - \bar{X}_n & \cdots & X_{nN} - \bar{X}_n \end{bmatrix}$$

v) Next compute the matrix:

$$X_{\text{new}} = FX.$$

How does PCA work?

Step 1: Standardize the data.
Step 2: Calculate the covariance matrix.
Step 3: Compute the eigenvectors and eigenvalues.
Step 4: Select the principal components.
Step 5: Project data onto the new basis.



Principle Component Analysis – Solved Example

- Given the data in Table, reduce the dimension from 2 to 1 using the Principal Component Analysis (PCA) algorithm.

Feature	Example 1	Example 2	Example 3	Example 4
X_1	4	8	13	7
X_2	11	4	5	14

Principle Component Analysis – Solved Example

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

Step 1: Calculate Mean

$$\bar{X}_1 = \tfrac{1}{4}(4+8+13+7) = 8,$$

$$\bar{X}_2 = \tfrac{1}{4}(11+4+5+14) = 8.5.$$

Principle Component Analysis – Solved Example

Step 2: Calculation of the covariance matrix.

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

Principle Component Analysis – Solved Example

Step 2: Calculation of the covariance matrix.

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

Principle Component Analysis – Solved Example

Step 2: Calculation of the covariance matrix.

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$\mathrm{Cov}(X_1, X_1) = \frac{1}{N-1} \sum_{k=1}^{N} (X_{1k} - \bar{X}_1)(X_{1k} - \bar{X}_1)$$

$$= \tfrac{1}{3}\left((4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2\right)$$

$$= 14$$

Principle Component Analysis – Solved Example

Step 2: Calculation of the covariance matrix.

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$\mathrm{Cov}(X_1, X_2) = \frac{1}{N-1} \sum_{k=1}^{N} (X_{1k} - \bar{X}_1)(X_{2k} - \bar{X}_2)$$

$$= \tfrac{1}{3}((4-8)(11-8.5) + (8-8)(4-8.5)$$

$$+ (13-8)(5-8.5) + (7-8)(14-8.5)$$

$$= -11$$

Step 2: Calculation of the covariance matrix.

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$\mathrm{Cov}(X_2, X_1) = \mathrm{Cov}(X_1, X_2)$$
$$= -11$$

$$\mathrm{Cov}(X_2, X_2) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{2k} - \bar{X}_2)(X_{2k} - \bar{X}_2)$$
$$= \frac{1}{3}\left((11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2\right)$$
$$= 23$$

Covariance matrix

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

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

Principle Component Analysis – Solved Example

Step 3: Eigenvalues of the covariance matrix

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

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

The characteristic equation of the covariance matrix is,

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

$$= \begin{vmatrix} 14 - \lambda & -11 \\ -11 & 23 - \lambda \end{vmatrix}$$

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

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

$$
\begin{aligned}
\lambda &= \tfrac{1}{2}(37 \pm \sqrt{565}) \\
&= 30.3849,\ 6.6151 \\
&= \lambda_1,\ \lambda_2 \quad \text{(say)}
\end{aligned}
$$

Principle Component Analysis – Solved Example

Step 4: Computation of the eigenvectors

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda I)\, U$$

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

$$= \begin{bmatrix} (14-\lambda)u_1 - 11u_2 \\ -11u_1 + (23-\lambda)u_2 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

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

$$(14-\lambda)u_1 - 11u_2 = 0$$

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

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

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 4: Computation of the eigenvectors

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

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

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

$$u_1 = 11t, \quad u_2 = (14 - \lambda\ )t$$

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

When t=1

$$U_1 = \begin{bmatrix} 11 \\ 14 - \lambda \end{bmatrix}.$$

Principle Component Analysis – Solved Example

Step 4: Computation of the eigenvectors

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$U_1 = \begin{bmatrix} 11 \\ 14 - \lambda_1 \end{bmatrix}.$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

- To find a unit eigenvector, we compute the length of U_1 which is given by,

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

$$\|U_1\| = \sqrt{11^2 + (14 - \lambda_1)^2}$$
$$= \sqrt{11^2 + (14 - 30.3849)^2}$$
$$= 19.7348$$

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Step 4: Computation of the eigenvectors

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$U_1 = \begin{bmatrix} 11 \\ 14 - \lambda_1 \end{bmatrix}.$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

- To find a unit eigenvector, we compute the length of U_1 which is given by,

$$\|U_1\| = \sqrt{11^2 + (14 - \lambda_1)^2}$$

$$= \sqrt{11^2 + (14 - 30.3849)^2}$$

$$= 19.7348$$

$$e_1 = \begin{bmatrix} 11/\|U_1\| \\ (14 - \lambda_1)/\|U_1\| \end{bmatrix}$$

$$= \begin{bmatrix} 11/19.7348 \\ (14 - 30.3849)/19.7348 \end{bmatrix}$$

e_1 $$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

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

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 4: Computation of the eigenvectors

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$U_1 = \begin{bmatrix} 11 \\ 14 - \lambda_1 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

- To find a unit eigenvector, we compute the length of U_1 which is given by,

$$\|U_1\| = \sqrt{11^2 + (14 - \lambda_1)^2}$$
$$= \sqrt{11^2 + (14 - 30.3849)^2}$$
$$= 19.7348$$

$$e_1 = \begin{bmatrix} 11/\|U_1\| \\ (14 - \lambda_1)/\|U_1\| \end{bmatrix} = \begin{bmatrix} 11/19.7348 \\ (14 - 30.3849)/19.7348 \end{bmatrix} = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

e_1

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

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

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 5: Computation of first principal components

$$e_1^T \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix}$$

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \quad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 5: Computation of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$e_1^T \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix}$$

$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix} \qquad \overline{X_1} = 8 \qquad \overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \qquad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\lambda_1 = 30.3849 \qquad \lambda_2 = 6.6151$$

$$e_1^T \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} X_{11} - \bar{X}_1 \\ X_{21} - \bar{X}_2 \end{bmatrix}$$

$$= 0.5574(X_{11} - \bar{X}_1) - 0.8303(X_{21} - \bar{X}_2)$$

$$= 0.5574(4 - 8) - 0.8303(11 - 8,5)$$

$$= -4.30535$$

Principle Component Analysis – Solved Example

Step 5: Computation of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

Feature	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14
First Principle Components	-4.3052	3.7361	5.6928	-5.1238

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

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

$$\lambda_1 = 30.3849$$

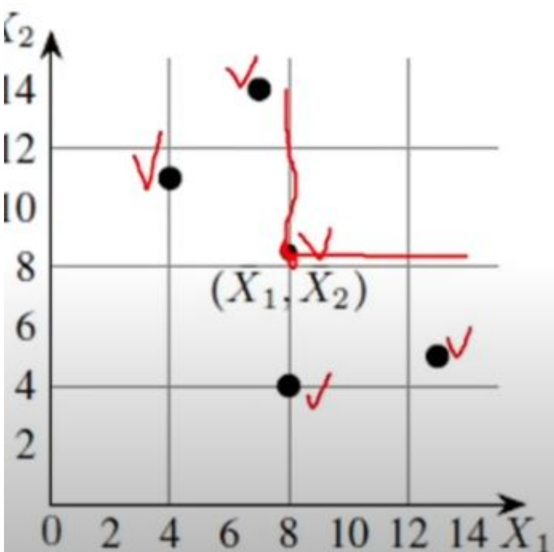
$$\lambda_2 = 6.6151$$

mean

Principle Component Analysis – Solved Example

Step 6: Geometrical meaning of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14



$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \quad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\lambda_1 = 30.3849$$

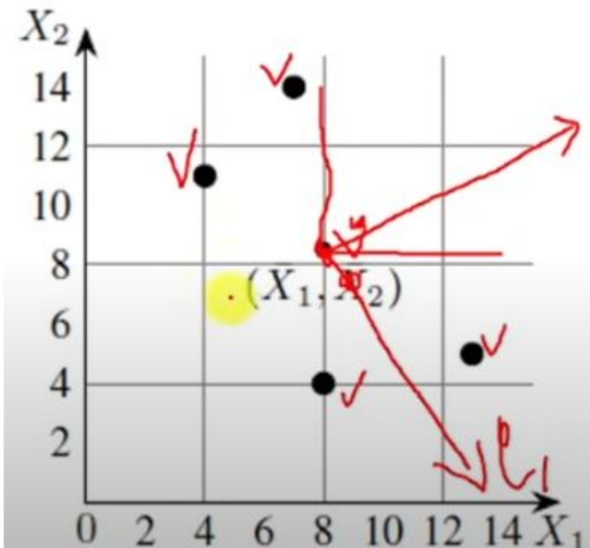
$$\lambda_2 = 6.6151$$

E1 and e2

Principle Component Analysis – Solved Example

Step 6: Geometrical meaning of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14



$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \quad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

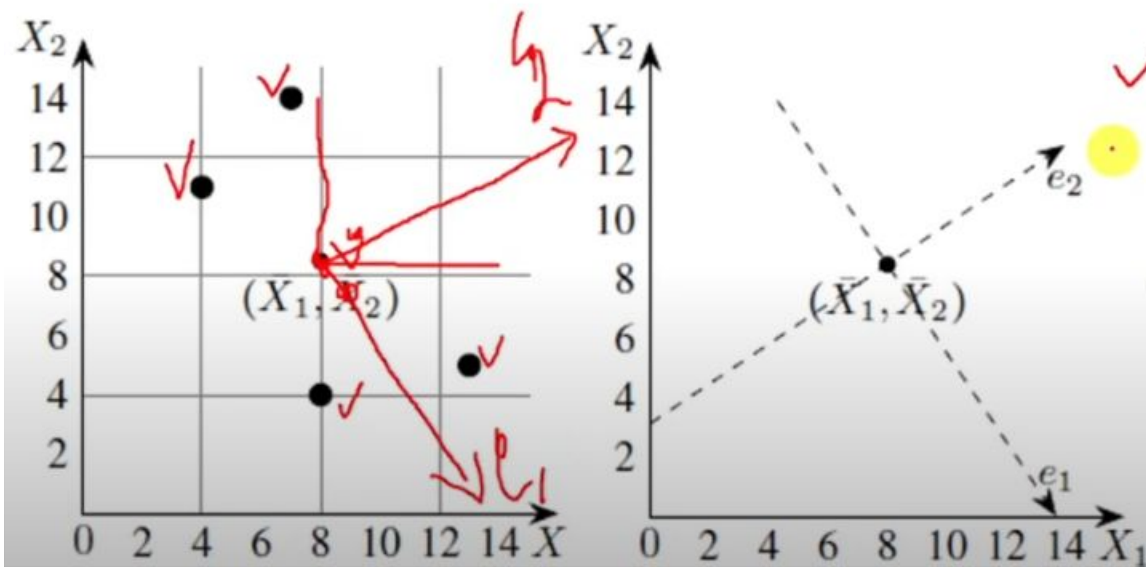
$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 6: Geometrical meaning of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14



$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \quad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

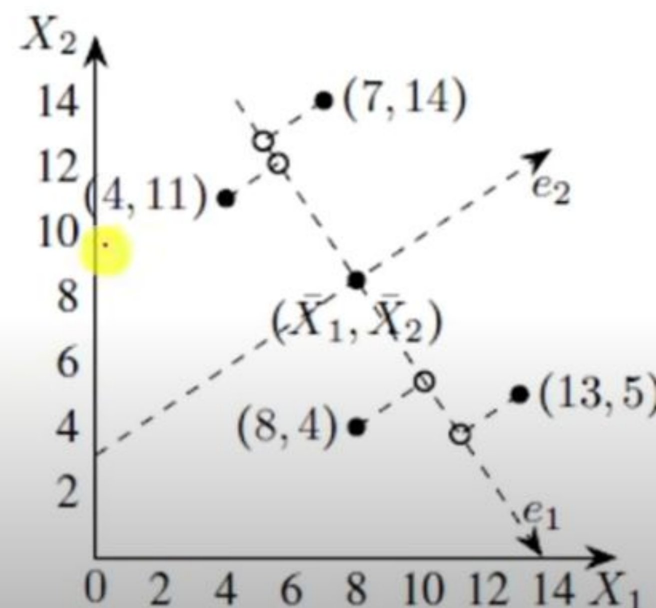
$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Project datapoint s to the first principal component

Principle Component Analysis – Solved Example

Step 6: Geometrical meaning of first principal components



F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\overline{X_1} = 8$$

$$\overline{X_2} = 8.5$$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix} \quad S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Principle Component Analysis – Solved Example

Step 5: Computation of first principal components

F	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14

Feature	Ex 1	Ex 2	Ex 3	Ex 4
X_1	4	8	13	7
X_2	11	4	5	14
First Principle Components	-4.3052	3.7361	5.6928	-5.1238

$\overline{X_1} = 8$

$\overline{X_2} = 8.5$

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

$\lambda_1 = 30.3849$

$\lambda_2 = 6.6151$

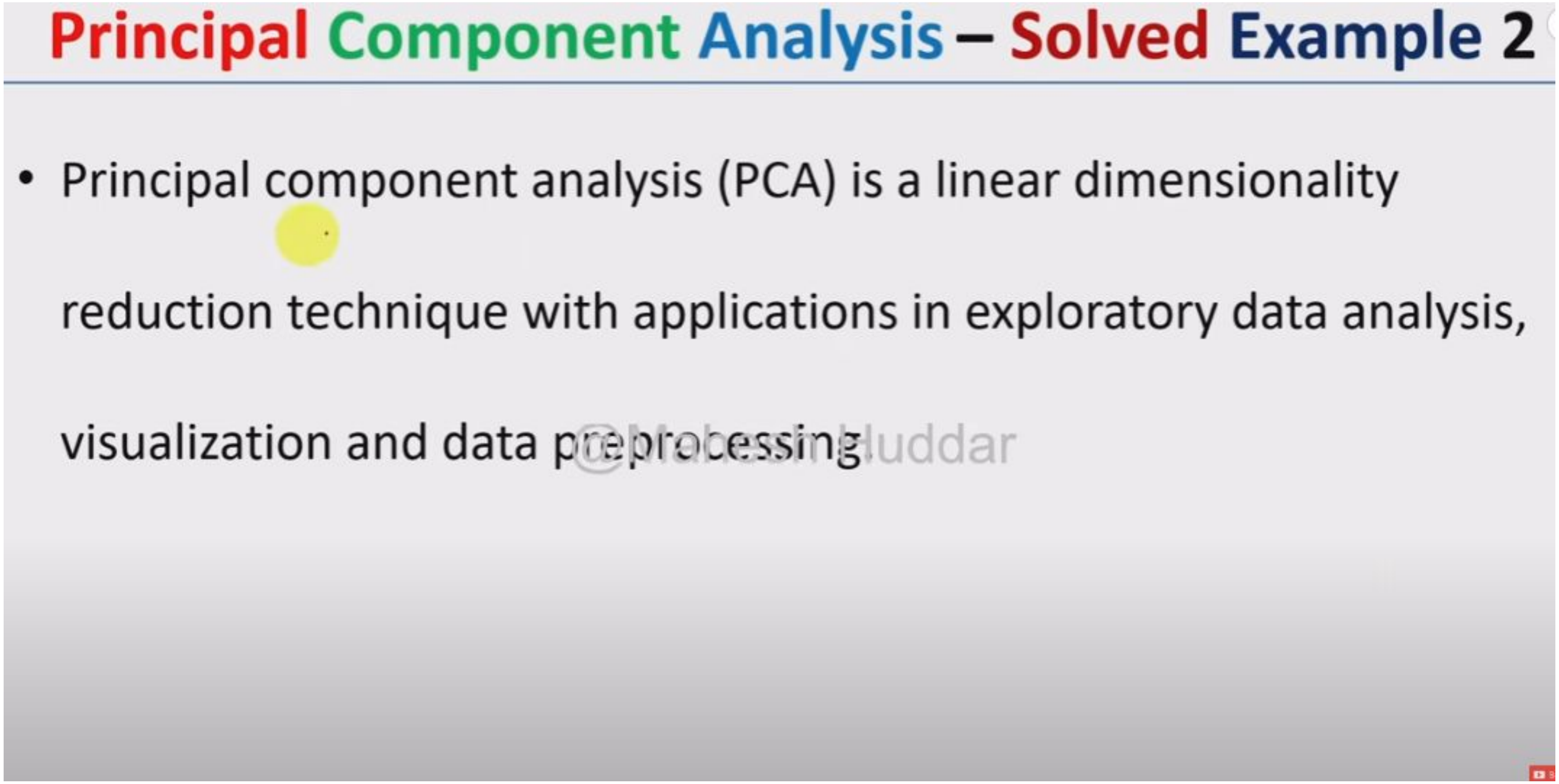
Example 2

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis

PCA Step by Step Solution

Solved Example #2


Principal Component Analysis – Solved Example 2
• Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.
@Mahesh Huddar

Principal Component Analysis – Solved Example 2

- Given the data in Table, reduce the dimension from 2 to 1 using the Principal Component Analysis (PCA) algorithm.

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 1: Compute Covariance Matrix

- $C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$
- $Cov(x,x) = \frac{\sum(x-\bar{x})(x-\bar{x})}{n-1}$
- $Cov(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 1: Compute Covariance Matrix

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$$

	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})*(Y-\bar{Y})$
	2.5	2.4	0.69	0.49	0.4761	0.2401	0.3381
	0.5	0.7	-1.31	-1.21	1.7161	1.4641	1.5851
	2.2	2.9	0.39	0.99	0.1521	0.9801	0.3861
	1.9	2.2	0.09	0.29	0.0081	0.0841	0.0261
	3.1	3.0	1.29	1.09	1.6641	1.1881	1.4061
	2.3	2.7	0.49	0.79	0.2401	0.6241	0.3871
	2	1.6	0.19	-0.31	0.0361	0.0961	-0.0589
	1	1.1	-0.81	-0.81	0.6561	0.6561	0.6561
	1.5	1.6	-0.31	-0.31	0.0961	0.0961	0.0961
	1.1	0.9	-0.71	-1.01	0.5041	1.0201	0.7171
Sum	18.1	19.1		$\rightarrow$	5.549	6.449	5.539
Mean	1.81	1.91		Sum/(n-1)	0.6166 ✓	0.7166 ✓	0.6154

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 1: Compute Covariance Matrix

$$C = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 2: Compute Eigen Values

$$C = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$$|C - \lambda I| = 0$$

$$\left| \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} 0.6166 - \lambda & 0.6154 \\ 0.6154 & 0.7166 - \lambda \end{bmatrix} \right| = 0$$

$$(0.6166 - \lambda) * (0.7166 - \lambda) - 0.6154^2 = 0$$

$$\lambda^2 - 1.333\lambda + 0.0630 = 0$$

$$\boldsymbol{\lambda_1 = 0.049}$$

$$\boldsymbol{\lambda_2 = 1.284}$$

Principal Component Analysis – Solved Example 2

Step 3: Compute Eigen Vectors $C = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$ $\lambda_1 = 0.049$ $\lambda_2 = 1.284$

- $Cv_1 = \lambda v_1$

$\sqrt{a^2+b^2}$

$v_1 = \begin{bmatrix} -1.084 \\ 1 \end{bmatrix}$

- $\begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} * \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.049 * \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$

- $0.6166x_1 + 0.6154y_1 = 0.0490x_1 \dots \dots (1)$
- $0.6154x_1 + 0.7166y_1 = 0.0490y_1 \dots \dots (2)$
- $0.5676x_1 = -0.6154y_1 \dots \dots (3)$
- $0.6154x_1 = -0.6676y_1 \dots \dots (4)$

$$v_1 = \begin{bmatrix} \dfrac{-1.084}{\sqrt{(-1.084)^2 + 1^2}} \\ \dfrac{1}{\sqrt{(-1.084)^2 + 1^2}} \end{bmatrix}$$

$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 3: Compute Eigen Vectors $C = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$ $\lambda_1 = 0.049$ $\lambda_2 = 1.284$

- $Cv_2 = \lambda v_2$
- $\begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} * \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 1.284 * \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$
- $0.6166x_2 + 0.6154y_2 = 1.284x_2 \ldots\ldots (5)$
- $0.6154x_2 + 0.7166y_2 = 1.284y_2 \ldots\ldots (6)$
- $0.6674x_2 = 0.6154y_2 \ldots\ldots (7)$
- $0.6154x_2 = 0.5674y_2 \ldots\ldots (8)$

$$v_2 = \begin{bmatrix} 0.922 \\ 1 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} \frac{0.922}{\sqrt{0.922^2 + 1^2}} \\ \frac{1}{\sqrt{0.922^2 + 1^2}} \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 4: Recast The Data Along The Principal Components Axes 3 2

$\underline{New} = \underline{F} * \underline{v_2}$

$$New = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \\ 1.9 & 2.2 \\ 3.1 & 3 \\ 2.3 & 2.7 \\ 2 & 1.6 \\ 1 & 1.1 \\ 1.5 & 1.6 \\ 1.1 & 0.9 \end{bmatrix} * \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

$\lambda_1 = 0.049$

$\lambda_2 = 1.284$

$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$

$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 4: Recast The Data Along The Principal Components Axes

$$New = F * v_2$$

$$New = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \\ 1.9 & 2.2 \\ 3.1 & 3 \\ 2.3 & 2.7 \\ 2 & 1.6 \\ 1 & 1.1 \\ 1.5 & 1.6 \\ 1.1 & 0.9 \end{bmatrix} * \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix} = \begin{bmatrix} 3.4565 \\ 0.853 \\ 3.6209 \\ 2.9033 \\ 4.3037 \\ 3.5416 \\ 2.53 \\ 1.4855 \\ 2.1915 \\ 1.4062 \end{bmatrix}$$

$$\lambda_1 = 0.049$$

$$\lambda_2 = 1.284$$

$$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

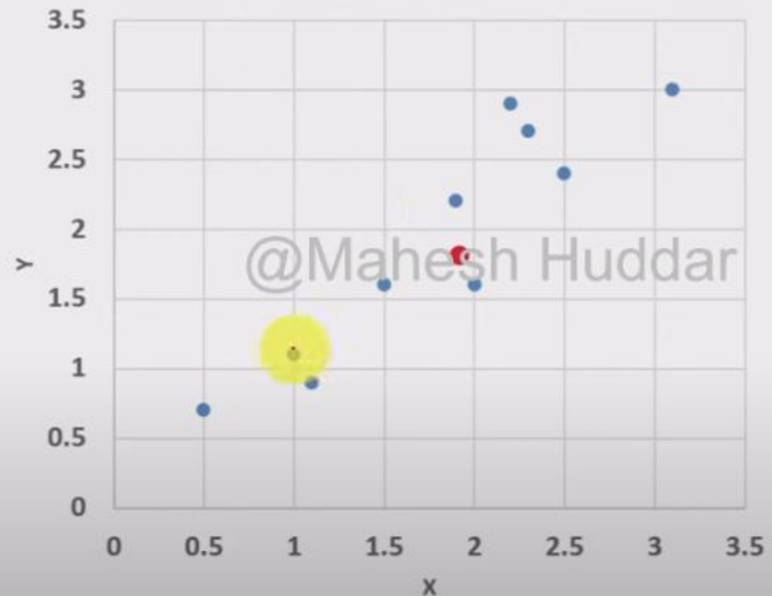
$$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 4: Recast The Data Along The Principal Components Axes

$$New = \begin{bmatrix} 3.4565 \\ 0.853 \\ 3.6209 \\ 2.9033 \\ 4.3037 \\ 3.5416 \\ 2.53 \\ 1.4855 \\ 2.1915 \\ 1.4062 \end{bmatrix}$$



$$\bar{x} = 1.91$$

$$\bar{y} = 1.81$$

$$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

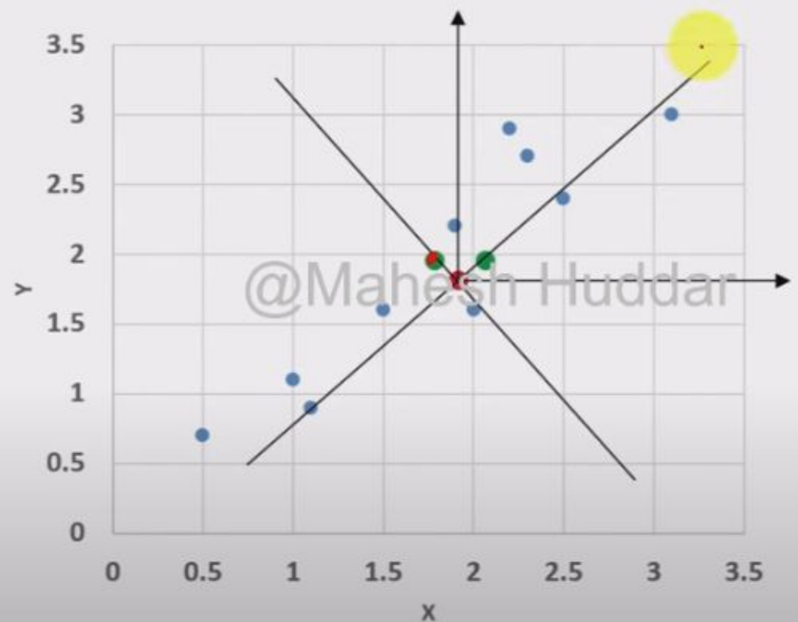
$$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 4: Recast The Data Along The Principal Components Axes

$$New = \begin{bmatrix} 3.4565 \\ 0.853 \\ 3.6209 \\ 2.9033 \\ 4.3037 \\ 3.5416 \\ 2.53 \\ 1.4855 \\ 2.1915 \\ 1.4062 \end{bmatrix}$$



$$\bar{x} = 1.91$$

$$\bar{y} = 1.81$$

$$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

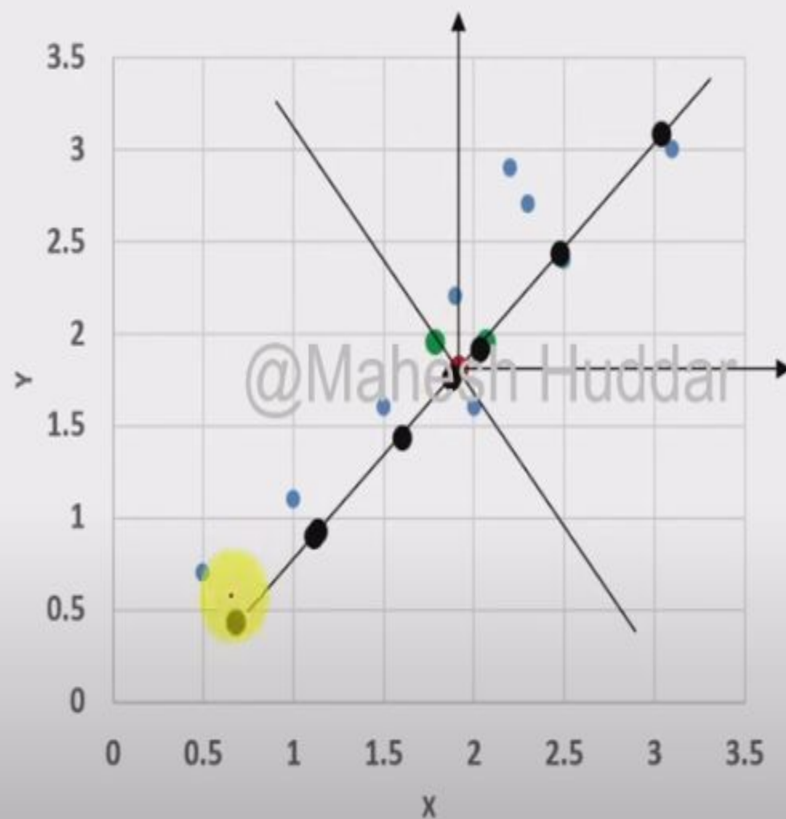
$$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Principal Component Analysis – Solved Example 2

Step 4: Recast The Data Along The Principal Components Axes

$$New = \begin{bmatrix} 3.4565 \\ 0.853 \\ 3.6209 \\ 2.9033 \\ 4.3037 \\ 3.5416 \\ 2.53 \\ 1.4855 \\ 2.1915 \\ 1.4062 \end{bmatrix}$$



$$\bar{x} = 1.91$$

$$\bar{y} = 1.81$$

$$v_1 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Applications of PCA

- Netflix Movie Recommendations
- Grocery Shopping
- Fitness Trackers
- Car Shopping
- Real Estate
- Manufacturing and Quality Control
- Sports Analytics
- Renewable Energy
- Smart Cities

Advantages of PCA

- Prevents Overfitting
- Speeds Up Other Machine Learning Algorithms
- Improves Visualization
- Dimensionality Reduction
- Noise Reduction

Limitations of PCA

- Linearity Assumption
- Loss of Interpretability
- Loss of Information
- Sensitivity to Scaling
- Orthogonal Components