# Case study 2

1. Customer Segmentation using PCA :

$$Data = \begin{bmatrix} 120 & 80 & 50 & 200 & 100 \\ 90 & 120 & 100 & 80 & 50 \\ 200 & 50 & 30 & 150 & 60 \\ 40 & 130 & 120 & 60 & 80 \end{bmatrix}$$

Using PCA to reduce dimensionality of the data while maintaining as much variance as possible.

* Standardizing the Data -

$$Z = \frac{X - \mu}{\sigma}$$

$X \equiv$ original value
$\mu \equiv$ mean
$\sigma \equiv$ standard deviation of product

1. Mean

$$\mu_{Electronics} = \frac{120+90+200+40}{4} = 112.5$$

$$\mu_{clothing} = \frac{80+120+50+130}{4} = 95$$

$$\mu_{Groceries} = \frac{50+100+30+120}{4} = 75$$

$$\mu_{Furniture} = \frac{100+150+60+180}{4} = 122.5$$

$$\mu_{Toys} = \frac{100+150+60+180}{4} = 72.5$$

mean vector, $\mu = [112.5, 95, 75, 122.5, 72.5]$

2. Standard deviation:

$$\sigma_{Electronics} = \sqrt{\frac{(120-112.5)^2+(90-112.5)^2+(200-112.5)^2+(40-112.5)^2}{4-1}}$$

$$= 68.95$$

$$\sigma_{clothing} = \sqrt{\frac{(80-95)^2+(120-95)^2+(50-95)^2+(130-95)^2}{4-1}}$$

$$= 36.74$$

$$\sigma_{Groceries} = \sqrt{\frac{(60-75)^2+(100-75)^2+(30-75)^2+(120-75)^2}{4-1}}$$

$$= 39.68$$

$\sigma_{Furniture} = 64.55$

$\sigma_{Toys} = 21.65$

s.d. vector $\sigma = [68.95, 36.74, 39.68, 64.55, 21.65]$

3. Standardizing:

$$z = \frac{X - \mu}{\sigma}$$

C1

- $z_{11} = \frac{120 - 112.5}{68.95} = 0.11$
- $z_{12} = \frac{80 - 95}{36.74} = -0.41$
- $z_{13} = \frac{50 - 75}{39.68} = -0.63$
- $z_{14} = \frac{200 - 122.5}{64.55} = 1.20$
- $z_{15} = \frac{100 - 72.5}{21.65} = 1.27$

$$Z = \begin{bmatrix} 0.11 & -0.41 & -0.63 & 1.20 & 1.27 \\ -0.133 & 0.68 & 0.63 & -0.66 & -1.04 \\ 1.27 & -1.22 & -1.13 & 0.43 & -0.58 \\ -1.05 & 0.95 & 1.13 & -0.97 & 0.35 \end{bmatrix}$$

* Covariance Matrix:

$$C = \frac{1}{n-1} Z^T Z$$

$$Z^T = \begin{bmatrix} 0.11 & -0.33 & 1.27 & -1.05 \\ -0.41 & 0.68 & -1.22 & 0.95 \\ -0.63 & 0.63 & -1.13 & 1.13 \\ 1.20 & -0.66 & 0.43 & -0.97 \\ 1.27 & -1.04 & -0.58 & 0.35 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 3.0476 & -2.1684 & -2.3314 & 1.8774 & -1.5596 \\ -2.1684 & 2.6936 & 2.7328 & -2.1488 & 1.3808 \\ -2.3314 & 2.7328 & 3.2042 & -2.2822 & 1.8472 \\ 1.8774 & -2.1488 & -2.2822 & 2.3598 & -1.3142 \\ -1.5596 & 1.3808 & 1.8472 & -1.3142 & 1.3318 \end{bmatrix}$$

Divide by $(n-1)$, $n = 4$

$$C = \frac{1}{3} \begin{bmatrix} 3.0476 & -2.1684 & -2.3314 & 1.8774 & -1.5596 \\ -2.1684 & 2.6936 & 2.7328 & -2.1488 & 1.3808 \\ -2.3314 & 2.7328 & 3.2042 & -2.2822 & 1.8472 \\ 1.8774 & -2.1488 & -2.2822 & 2.3598 & -1.3142 \\ -1.5596 & 1.3808 & 1.8472 & -1.3142 & 1.3318 \end{bmatrix}$$

$$C = \begin{bmatrix} 1.0159 & -0.1288 & -0.7711 & 0.6258 & -0.5199 \\ -0.7228 & 0.8979 & 0.9109 & -0.7163 & 0.4603 \\ -0.7771 & 0.9109 & 1.0681 & -0.7607 & 0.6157 \\ 0.6258 & -0.7163 & -0.7607 & 0.7866 & -0.4381 \\ -0.5199 & 0.4603 & 0.6157 & -0.4381 & 0.4439 \end{bmatrix}$$

- Find Eigenvalues

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

$$\lambda = [3.5524,\ 0.3287,\ 0.1820,\ 0.1399,\ 0.0094]$$

- Eigenvectors corresponding to each column are:

$$\begin{bmatrix} 0.4661 & -0.8483 & 0.1858 & 0.0935 & 0.1416 \\ -0.4765 & -0.2864 & -0.1800 & -0.6323 & 0.5086 \\ -0.5277 & -0.2938 & 0.4994 & -0.1039 & -0.6124 \\ 0.4236 & 0.3193 & 0.6196 & -0.5724 & 0.0842 \\ -0.3127 & 0.1003 & 0.5475 & 0.5030 & 0.5825 \end{bmatrix}$$

* Select Principal components:

$$\lambda_1 = 3.5524 \quad , \quad \lambda_2 = 0.3287$$

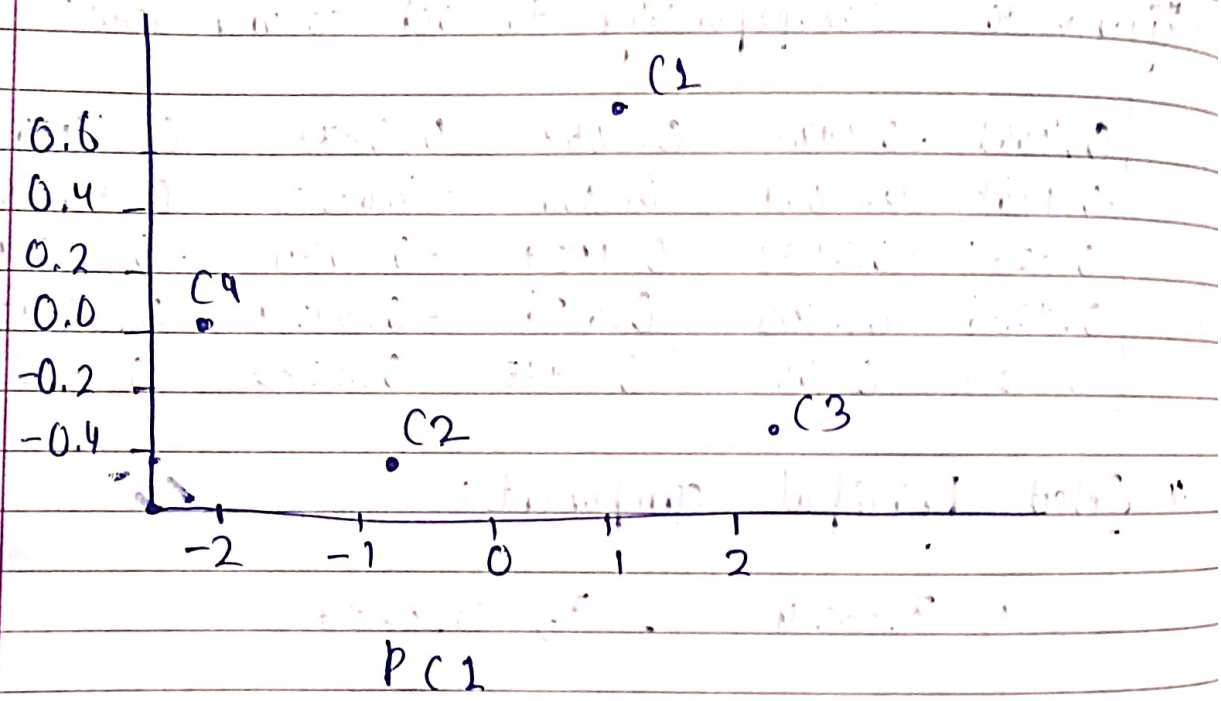
Eigenvectors corresponding to these eigenvalues

$$W = \begin{bmatrix} 0.4661 & -0.8483 \\ -0.4765 & -0.2864 \\ -0.5277 & -0.2938 \\ 0.4236 & 0.3193 \\ -0.3127 & 0.1003 \end{bmatrix}$$

$$Z.W = \begin{bmatrix} 0.6902 & -0.7198 \\ -0.7646 & -0.4156 \\ 2.1331 & -0.3169 \\ -2.0581 & 0.0121 \end{bmatrix} \begin{matrix} C1 \\ C2 \\ C3 \\ C4 \end{matrix}$$

PC1: largest amount of variance in purchasing behaviour.

PC2: second amount differentiating customer behaviour but explains less variance.




- C2 and C4 are close, therefore, they have similar purchasing behaviours.
- C1 and C3 are far away, indicating distinct purchasing behaviour.

# Predicting House Prices Using Linear Regression with Matrix Operations -

| | Size | Rooms | Distance | Price |
|---|---|---|---|---|
| Data = | 1500 | 3 | 5 | 400,000 |
| | 2000 | 4 | 10 | 500,000 |
| | 1200 | 2 | 3 | 350,000 |
| | 1800 | 3 | 8 | 450,000 |

$Y$ = target variable (house prices)
$X$ = matrix of input features
$\epsilon$ = error term
$\beta$ = vector of coefficients

$$Y = X \cdot \beta + \epsilon$$

$$\beta = (X^T X)^{-1} X^T Y$$

* $$X = \begin{bmatrix} 1 & 1500 & 3 & 5 \\ 1 & 2000 & 4 & 10 \\ 1 & 1200 & 2 & 3 \\ 1 & 1800 & 3 & 8 \end{bmatrix}$$

$$Y = \begin{bmatrix} 400000 \\ 500000 \\ 350000 \\ 450000 \end{bmatrix}$$

* $X^T X$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1500 & 2000 & 1200 & 1800 \\ 3 & 4 & 2 & 3 \\ 5 & 10 & 3 & 8 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1500 & 2000 & 1200 & 1800 \\ 3 & 4 & 2 & 3 \\ 5 & 10 & 3 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1500 & 3 & 5 \\ 1 & 2000 & 4 & 10 \\ 1 & 1200 & 2 & 3 \\ 1 & 1800 & 3 & 8 \end{bmatrix}$$

* $X^T Y$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1500 & 2000 & 1200 & 1800 \\ 3 & 4 & 2 & 3 \\ 5 & 10 & 3 & 8 \end{bmatrix} \begin{bmatrix} 400000 \\ 500000 \\ 350000 \\ 450000 \end{bmatrix}$$

* $X^T X := \begin{bmatrix} 4 & 6500 & 12 & 26 \\ 6500 & 10930000 & 20300 & 45500 \\ 12 & 20300 & 38 & 85 \\ 26 & 45500 & 85 & 198 \end{bmatrix}$

$$X^T Y := \begin{bmatrix} 1700000 \\ 2830000000 \\ 5250000 \\ 11650000 \end{bmatrix}$$

Page No.:
Date:

* calculate coefficients $\beta$ for the linear regression model are:

$$\beta = \begin{bmatrix} 266666.67 \\ 0 \\ 16666.67 \\ 16666.67 \end{bmatrix}$$

- intercept term is approx. 266,666.67.
- coefficient for Size (sq ft) is 0.
- coefficient for rooms is 16,666.67.
- distance to city center is 16,666.67.

which suggests that Size feature may not contribute to deciding Price, rooms & distance have more impact

3. Collaborative Filtering for Movie Recommendations Using Matrix Factorization -

Data -

| User ID | Movie 1 Rating | Movie 2 Rating | Movie 3 Rating | Movie 4 Rating |
|---|---|---|---|---|
| U1 | 5 | 3 | 0 | 1 |
| U2 | 4 | 0 | 5 | 1 |
| U3 | 0 | 2 | 4 | 3 |
| U4 | 3 | 5 | 2 | 0 |

$$(R)matrix = \begin{bmatrix} 5 & 3 & 0 & 1 \\ 4 & 0 & 5 & 1 \\ 0 & 2 & 4 & 3 \\ 3 & 5 & 2 & 0 \end{bmatrix}$$

$$R = U \Sigma V^T$$

$$R^T = \begin{bmatrix} 5 & 4 & 0 & 3 \\ 3 & 0 & 2 & 5 \\ 0 & 5 & 4 & 2 \\ 1 & 1 & 3 & 0 \end{bmatrix}$$

$$R_1 = RR^T = \begin{bmatrix} 5 & 3 & 0 & 1 \\ 4 & 0 & 5 & 1 \\ 0 & 2 & 4 & 3 \\ 3 & 5 & 2 & 0 \end{bmatrix} \begin{bmatrix} 5 & 4 & 0 & 3 \\ 3 & 0 & 2 & 5 \\ 0 & 5 & 4 & 2 \\ 1 & 1 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 35 & 21 & 9 & 30 \\ 21 & 42 & 23 & 22 \\ 9 & 23 & 29 & 18 \\ 30 & 22 & 18 & 38 \end{bmatrix}$$

$$R_2 = R^TR = \begin{bmatrix} 5 & 4 & 0 & 3 \\ 3 & 0 & 2 & 5 \\ 0 & 5 & 4 & 2 \\ 1 & 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} 5 & 3 & 0 & 1 \\ 4 & 0 & 5 & 1 \\ 0 & 2 & 4 & 3 \\ 3 & 5 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 50 & 30 & 26 & 9 \\ 30 & 38 & 18 & 9 \\ 36 & 18 & 45 & 17 \\ 9 & 9 & 17 & 11 \end{bmatrix}$$

* Find value of left singular vector U.
* Find eigen values and eigenvector of $A \times A^T$.
* Find value of $V^T$
* Find eigen values and eigenvectors of $A^T \times A$.
* Singular values are square root of eigenvalues and are sorted in descending order.

$$|R_1 - \lambda I| = 0$$

$$\Rightarrow \begin{bmatrix} 35 & 21 & 9 & 30 \\ 21 & 42 & 23 & 22 \\ 9 & 23 & 29 & 18 \\ 30 & 22 & 18 & 38 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 0$$

$$U = \begin{bmatrix} -0.493 & 0.563 & 0.231 & -0.618 \\ -0.549 & -0.474 & 0.639 & 0.252 \\ -0.385 & -0.569 & -0.581 & -0.435 \\ -0.553 & 0.365 & -0.442 & 0.603 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 9.956 & 0 & 0 & 0 \\ 0 & 5.262 & 0 & 0 \\ 0 & 0 & 3.701 & 0 \\ 0 & 0 & 0 & 1.866 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.635 & -0.503 & -0.542 & -0.220 \\ 0.383 & 0.452 & -0.744 & -0.307 \\ 0.654 & -0.719 & -0.002 & -0.243 \\ -0.146 & 0.156 & 0.390 & -0.895 \end{bmatrix}$$

- U contains latent user factors. Each row represents the preferences or characteristics of a user in the latent factor space.
- $\Sigma$: Indicates importance or significance of each latent factor.
- $V^T$: contains latent movie factors.

4. Dimensionality reduction in Image processing using (SVD):

$$(X)\,Data = \begin{bmatrix} 255 & 200 & 150 & 100 \\ 200 & 150 & 100 & 50 \\ 150 & 100 & 50 & 0 \\ 100 & 50 & 0 & 0 \end{bmatrix}$$

S.V.D. of $X = U \Sigma V^T$

$$X^T = \begin{bmatrix} 255 & 200 & 150 & 100 \\ 200 & 150 & 100 & 50 \\ 150 & 100 & 50 & 0 \\ 100 & 50 & 0 & 0 \end{bmatrix}$$

$$XX^T = \begin{bmatrix} 255 & 200 & 150 & 100 \\ 200 & 150 & 100 & 50 \\ 150 & 100 & 50 & 0 \\ 100 & 50 & 0 & 0 \end{bmatrix} \begin{bmatrix} 255 & 200 & 150 & 100 \\ 200 & 150 & 100 & 50 \\ 150 & 100 & 50 & 0 \\ 100 & 50 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 137525 & 101000 & 65750 & 35500 \\ 101000 & 75000 & 50000 & 27500 \\ 65750 & 50000 & 35000 & 20000 \\ 35500 & 27500 & 20000 & 12500 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 137525 & 101000 & 65750 & 35500 \\ 101000 & 75000 & 50000 & 27500 \\ 65750 & 50000 & 35000 & 20000 \\ 35500 & 27500 & 20000 & 12500 \end{bmatrix}$$

$$U: \begin{bmatrix} -0.493 & 0.563 & 0.231 & -0.618 \\ -0.549 & -0.414 & 0.639 & 0.252 \\ -0.385 & -0.569 & -0.581 & -0.435 \\ -0.553 & 0.365 & -0.442 & 0.603 \end{bmatrix}$$

$$\Sigma: \begin{bmatrix} 9.956 & 0 & 0 & 0 \\ 0 & 5.262 & 0 & 0 \\ 0 & 0 & 3.701 & 0 \\ 0 & 0 & 0 & 1.866 \end{bmatrix}$$

$$V^T: \begin{bmatrix} -0.635 & -0.503 & -0.542 & -0.220 \\ 0.383 & 0.452 & -0.744 & -0.307 \\ 0.654 & -0.719 & -0.002 & -0.234 \\ -0.146 & 0.15 & 0.390 & -0.895 \end{bmatrix}$$

- Now reconstruct the Image while preserving most important features.
- choose no. of singular values $k=2$.
- reconstruct the compressed Image.

Reconstructed image :

$$\begin{bmatrix} 255 & 200 & 150 & 100 \\ 200 & 150 & 100 & 50 \\ 150 & 100 & 50 & 2.75\times10^{-4} \\ 100 & 50 & 1.94\times10^{-14} & 1.61\times10^{-4} \end{bmatrix}$$

compressed image (with k=2) :

$$\begin{bmatrix} 253.503 & 201.373 & 153.093 & 96.142 \\ 201.373 & 148.561 & 97.665 & 53.108 \\ 153.093 & 97.665 & 42.167 & 9.206 \\ 96.142 & 53.108 & 9.206 & -11.000 \end{bmatrix}$$

Date:

## 5. Network Flow Analysis Using Eigenvalues and Eigenvectors:

Data: adjacency matrix representing network traffic b/w servers:

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

* Eigenvalue Decomposition

1. Calculate eigenvalues and eigenvectors of adjacency matrix

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

$$\det\left(\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right) = 0$$

Eigenvalues: $[2.561 \quad -5.003 \times 10^{-17} \quad -1.561 \quad -1.000]$

Eigen vectors:

$$\begin{bmatrix} -.435 & -.707 & .557 & 1.570 \times 10^{-16} \\ -.557 & -1.5 \times 10^{-16} & -.435 & -.707 \\ -.557 & -1.89 \times 10^{-16} & -.435 & -.707 \\ -.435 & .707 & .557 & 1.57 \times 10^{-16} \end{bmatrix}$$

sorted eigen values:

$$\begin{bmatrix} 2.561 & -5.003 \times 10^{-17} & -1. & -1.561 \end{bmatrix}$$

- Eigen values indicate how strongly the network is connected along respective eigenvectors.

  Higher eigen values = important nodes

- largest eigenvalues represent the centrality of the nodes.

- By analyzing eigenvectors we can identify which nodes (servers) are more central or influential in network.

Page No.:
Date:

## 6. Topic Modeling with LSA Using SVD:

### Data:

Term document matrix for 4 documents and 5 terms:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

### Applying SVD

$$A = U \Sigma V^T$$

- $U$ : left singular vectors (terms by topics).
- $\Sigma$ : singular values.
- $V^T$ : right singular vectors (topics by documents)

* $\Sigma$: singular values, larger = more important

U: columns represent the latent semantic topics, and each term is associated with a vector indicating its importance for each topic.

$V^T$: The rows represent the topic-document associations, and each document is associated with a vector representing its relationship with the topics.

Reconstructed matrix: use reduced matrices to reconstruct an approx. of the original term-document matrix with reduced dimensions.

- columns of U represent important terms for each topic.
- rows of V show how document is related to the discovered topics.

- $\Sigma$: show importance of the topics.
  first two topics capture most variance in the data.
- Topic identification: columns of $U_k$ show importance of each term of the topic.

('Topic1' has more common terms, 'Topic2' has more different terms)

Page No.:
Date:

- Document-Topic associations: rows of $V_k^T$ represent how each document relates to the topics.

## 7. Stock Price Prediction Using Linear Algebra & PCA:

Data:

| Stock A | Stock B | Stock C | Stock D | Price |
|---|---|---|---|---|
| 120 | 85 | 95 | 150 | 100 |
| 110 | 80 | 90 | 140 | 90 |
| 130 | 90 | 100 | 160 | 110 |
| 115 | 82 | 93 | 145 | 95 |

* Apply PCA for Dimensionality reduction -

- Using PCA to reduce dimensionality while preserving as much variability as possible.
- identification of principal components that capture the most variance in the data.
- projection of data onto these components to represent the data in a lower-dimensional space.

$$\text{Data} = \begin{bmatrix} 120 & 85 & 95 & 150 \\ 110 & 80 & 90 & 140 \\ 130 & 90 & 100 & 160 \\ 115 & 82 & 93 & 145 \end{bmatrix}$$

Eigenvectors:

$$\begin{bmatrix} 0.632 & 0.321 & 0.316 & 0.632 \\ 0.004 & -0.704 & 0.709 & 0.004 \end{bmatrix}$$

Transformed data (in PCA space):

$$\begin{bmatrix} 1.977 & -0.161 \\ -13.833 & -0.282 \\ 17.78 & -0.040 \\ -5.93 & 0.484 \end{bmatrix}$$

* Using PCA transformed data for Linear regression

Y: target variable: stock prices

$$Y = X\beta + \epsilon$$

Target → Y

Input → X

Coefficient → β

error term (intercept) → ε

Coefficients: [ 0.63243692 0.00485163]

Intercept: 98.75

Page No.:
Date:
YOUVA

Conclusion:

- MSE: A low MSE value indicates that the predictions close to the actual values.
- R-squared: A high $R^2$ means model is explaining most of variance in the stock prices.
- We used PCA to reduce the dimensionality of the stock data and then trained a linear regression model to predict stock prices.

8. Face recognition using Eigen-faces:

Process:

- Convert facial images to Grayscale Matrices-
  - each pixel's intensity in a grayscale image becomes a numerical value.
  - image (m x n), flatten it into a vector of length (m x n).
- Data Matrix -
  - for N face images, create a data matrix X where each column is a vectorized image.
- Center the Data -
  - calculate mean of ($\mu$)
  - subtract the mean face from each image:

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

- Eigenface / Eigenvalues calculation -
  - Covariance Matrix: $C = \frac{1}{N} X^T_{CENTERED} X_{CENTERED}$.
  - find eigenvalues ($\lambda$) and eigenvectors of C.
  - sort eigenvalues in descending order and keep the top k eigenvectors.

- Projection:
  - $X_{new}$, $X_{new\,centered} = X_{new} - \mu$.
  - Project it onto the eigenface space: projection $= V_k^T X_{new\,centered}$
    $V_k$ contains the top k eigenfaces.
- Recognition:
  - Project faces onto eigenface space.
  - compare projections:
    - compute the distance between the new face's projection and the projections of stored faces in the database.
    - Identify the closest match.

9. Climate Data Analysis Using PCA:

Data:

| Year | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| 2000 | 30 | 28 | 25 | 32 |
| 2001 | 31 | 29 | 26 | 33 |
| 2002 | 30 | 28 | 25 | 32 |
| 2003 | 32 | 30 | 27 | 34 |

Apply PCA:

- $A = \begin{bmatrix} 30 & 28 & 25 & 32 \\ 31 & 29 & 26 & 33 \\ 30 & 28 & 25 & 32 \\ 32 & 30 & 27 & 34 \end{bmatrix}$
- Center the Data ($X_{centered}$):
- Use the centered data matrix to calculate the covariance matrix.
- Perform Eigenvalue Decomposition:
  - Find eigenvalues and eigenvectors of covariance matrix.
  - sort it into descending order.
- Select Principal components:
- Project data onto Principal components:

Page No.:
Date:

* PCA helps identify the main factors influencing the data across stations.
It simplifies dataset while preserving critical information, enabling insights into trends, correlations or differences between stations over time.