Understanding Variance, Covariance, and Centralized Data Matrices

February 19, 2025

Overview

What is a Centralized Data Matrix?

What is a Centralized Data Matrix?

- A centralized data matrix is obtained by subtracting the mean from each data point.

What is a Centralized Data Matrix?

- A centralized data matrix is obtained by subtracting the mean from each data point.
- It centers the data around the origin, making it easier to analyze variance and covariance.

What is a Centralized Data Matrix?

- A centralized data matrix is obtained by subtracting the mean from each data point.
- It centers the data around the origin, making it easier to analyze variance and covariance.

What is a Centralized Data Matrix?

- A centralized data matrix is obtained by subtracting the mean from each data point.
- It centers the data around the origin, making it easier to analyze variance and covariance.

Formula:

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

The centralized data matrix serves several important purposes in data analysis and machine learning:

- **Remove Bias:** By centering the data around the mean, you eliminate bias that might affect the analysis. This helps in understanding the underlying structure of the data without the influence of varying means.

The centralized data matrix serves several important purposes in data analysis and machine learning:

- **Remove Bias:** By centering the data around the mean, you eliminate bias that might affect the analysis. This helps in understanding the underlying structure of the data without the influence of varying means.
- **Facilitate PCA:** In Principal Component Analysis (PCA), centralization is a crucial step. PCA requires data to be centered to properly identify the directions of maximum variance, leading to meaningful principal components.

The centralized data matrix serves several important purposes in data analysis and machine learning:

- **Remove Bias:** By centering the data around the mean, you eliminate bias that might affect the analysis. This helps in understanding the underlying structure of the data without the influence of varying means.
- **Facilitate PCA:** In Principal Component Analysis (PCA), centralization is a crucial step. PCA requires data to be centered to properly identify the directions of maximum variance, leading to meaningful principal components.
- **Simplify Interpretation:** Centering the data makes the interpretation of results clearer. For example, the principal components represent directions of variance relative to the mean rather than absolute values.

The centralized data matrix serves several important purposes in data analysis and machine learning:

- **Remove Bias:** By centering the data around the mean, you eliminate bias that might affect the analysis. This helps in understanding the underlying structure of the data without the influence of varying means.
- **Facilitate PCA:** In Principal Component Analysis (PCA), centralization is a crucial step. PCA requires data to be centered to properly identify the directions of maximum variance, leading to meaningful principal components.
- **Simplify Interpretation:** Centering the data makes the interpretation of results clearer. For example, the principal components represent directions of variance relative to the mean rather than absolute values.
- **Normalize Feature Contributions:** In models that assume normally distributed data, centralizing helps ensure that each feature contributes equally to the analysis, reducing the risk of one feature dominating due to its scale.

Example of Centralized Data Matrix

Consider the dataset:

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

Example of Centralized Data Matrix

Consider the dataset:

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

1. Calculate the mean:

Example of Centralized Data Matrix

Consider the dataset:

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

1. Calculate the mean:

$$\mu = \frac{2+4+6}{3} = 4$$

Example of Centralized Data Matrix

Consider the dataset:

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

1. Calculate the mean:

$$\mu = \frac{2+4+6}{3} = 4$$

2. Centralize the data:

Example of Centralized Data Matrix

Consider the dataset:

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

1. Calculate the mean:

$$\mu = \frac{2+4+6}{3} = 4$$

2. Centralize the data:

$$X_{centered} = \begin{bmatrix} 2-4 \\ 4-4 \\ 6-4 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Example

Consider the following dataset with three features:

$$X = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 6 \\ 5 & 8 & 8 \\ 6 & 10 & 10 \end{bmatrix}$$

Example

Consider the following dataset with three features:

$$X = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 6 \\ 5 & 8 & 8 \\ 6 & 10 & 10 \end{bmatrix}$$

Step 1: Calculate the Mean of Each Feature First, calculate the mean of each feature (column):

Example

Consider the following dataset with three features:

$$X = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 6 \\ 5 & 8 & 8 \\ 6 & 10 & 10 \end{bmatrix}$$

Step 1: Calculate the Mean of Each Feature First, calculate the mean of each feature (column):

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \frac{2+3+5+6}{4} \\ \frac{3+5+8+10}{4} \\ \frac{5+6+8+10}{4} \end{bmatrix} = \begin{bmatrix} 4 \\ 6.5 \\ 7.25 \end{bmatrix}$$

Step 2: Centralize the Data Next, subtract the mean vector from each row of the original dataset:

$$X_{centered} = \begin{bmatrix} 2-4 & 3-6.5 & 5-7.25 \\ 3-4 & 5-6.5 & 6-7.25 \\ 5-4 & 8-6.5 & 8-7.25 \\ 6-4 & 10-6.5 & 10-7.25 \end{bmatrix}$$

Calculating each entry:

$$X_{centered} = \begin{bmatrix} -2 & -3.5 & -2.25 \\ -1 & -1.5 & -1.25 \\ 1 & 1.5 & 0.75 \\ 2 & 3.5 & 2.75 \end{bmatrix}$$

What is Variance?

- Variance measures the spread of a set of numbers.

What is Variance?

- Variance measures the spread of a set of numbers.
- It quantifies how far each number in the dataset is from the mean.

What is Variance?

- Variance measures the spread of a set of numbers.
- It quantifies how far each number in the dataset is from the mean.

Formula:

$$Var(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Where:

- N: Number of observations
- μ: Mean of the dataset

Example of Variance

Consider the dataset: $X = \{2, 4, 4, 4, 5, 5, 7, 9\}$

Example of Variance

Consider the dataset: $X = \{2, 4, 4, 4, 5, 5, 7, 9\}$

1. Calculate the mean:

$$\mu = \frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5$$

Example of Variance

Consider the dataset: $X = \{2, 4, 4, 4, 5, 5, 7, 9\}$

1. Calculate the mean:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = 5$$

2. Calculate variance:

$$Var(X) = \frac{1}{8}\left((2-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2\right)$$

Example of Variance

Consider the dataset: $X = \{2, 4, 4, 4, 5, 5, 7, 9\}$

1. Calculate the mean:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = 5$$

2. Calculate variance:

$Var(X) = \frac{1}{8}\left((2-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(5-5)^2+(7-5)^2+(9-5)^2\right)$

3. Solve:

$$Var(X) = \frac{1}{8}(9+1+1+1+0+0+4+16) = \frac{32}{8} = 4$$

What is Covariance?

- Covariance measures how two variables change together.

What is Covariance?

- Covariance measures how two variables change together.
- Positive covariance indicates both variables tend to increase together.

What is Covariance?

- Covariance measures how two variables change together.
- Positive covariance indicates both variables tend to increase together.
- Negative covariance indicates one variable increases while the other decreases.

What is Covariance?

- Covariance measures how two variables change together.
- Positive covariance indicates both variables tend to increase together.
- Negative covariance indicates one variable increases while the other decreases.

What is Covariance?

- Covariance measures how two variables change together.
- Positive covariance indicates both variables tend to increase together.
- Negative covariance indicates one variable increases while the other decreases.

Formula:

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

Where:

- μ_X and μ_Y: Means of datasets X and Y

Example of Covariance

Consider the datasets:

$$X = \{2, 4, 6\}, \quad Y = \{1, 3, 5\}$$

Example of Covariance

Consider the datasets:

$$X = \{2, 4, 6\}, \quad Y = \{1, 3, 5\}$$

1. Calculate means:

$$\mu_X = \frac{2+4+6}{3} = 4, \quad \mu_Y = \frac{1+3+5}{3} = 3$$

Example of Covariance

Consider the datasets:

$$X = \{2, 4, 6\}, \quad Y = \{1, 3, 5\}$$

1. Calculate means:

$$\mu_X = \frac{2+4+6}{3} = 4, \quad \mu_Y = \frac{1+3+5}{3} = 3$$

2. Calculate covariance:

$$Cov(X, Y) = \frac{1}{3}((2-4)(1-3) + (4-4)(3-3) + (6-4)(5-3))$$

Example of Covariance

Consider the datasets:

$$X = \{2, 4, 6\}, \quad Y = \{1, 3, 5\}$$

1. Calculate means:

$$\mu_X = \frac{2+4+6}{3} = 4, \quad \mu_Y = \frac{1+3+5}{3} = 3$$

2. Calculate covariance:

$$Cov(X, Y) = \frac{1}{3}((2-4)(1-3) + (4-4)(3-3) + (6-4)(5-3))$$

3. Solve:

$$Cov(X, Y) = \frac{1}{3}((-2)(-2) + 0 + (2)(2)) = \frac{8}{3} \approx 2.67$$

Covariance in Matrix Form

- The covariance matrix is a square matrix

Covariance in Matrix Form

- The covariance matrix is a square matrix
- It provides a measure of the covariance between pairs of variables in a dataset.

Covariance in Matrix Form

- The covariance matrix is a square matrix
- It provides a measure of the covariance between pairs of variables in a dataset.
- Each element of the matrix represents the covariance between two variables.

Covariance in Matrix Form

- The covariance matrix is a square matrix
- It provides a measure of the covariance between pairs of variables in a dataset.
- Each element of the matrix represents the covariance between two variables.

For a dataset with n observations and m features (variables), the covariance matrix C is defined as:

$$C = \frac{1}{N-1} X^T X$$

Covariance in Matrix Form

- The covariance matrix is a square matrix
- It provides a measure of the covariance between pairs of variables in a dataset.
- Each element of the matrix represents the covariance between two variables.

For a dataset with n observations and m features (variables), the covariance matrix C is defined as:

$$C = \frac{1}{N-1} X^T X$$

where X is the centered data matrix, and N is the number of observations.

Properties of Covariance Matrix

- Symmetric: $C_{ij} = C_{ji}$

Properties of Covariance Matrix

- Symmetric: $C_{ij} = C_{ji}$
- Diagonal Elements: The diagonal elements represent the variance of each variable.

Properties of Covariance Matrix

- Symmetric: $C_{ij} = C_{ji}$
- Diagonal Elements: The diagonal elements represent the variance of each variable.
- Off-diagonal Elements: The off-diagonal elements represent the covariance between pairs of variables.

Properties of Covariance Matrix

- Symmetric: $C_{ij} = C_{ji}$
- Diagonal Elements: The diagonal elements represent the variance of each variable.
- Off-diagonal Elements: The off-diagonal elements represent the covariance between pairs of variables.
- Semi positive definite matrix

Example

Consider a dataset with two features, X and Y:

Example

Consider a dataset with two features, X and Y:

X	Y
2	3
3	5
5	7
8	10

Example

Consider a dataset with two features, X and Y:

X	Y
2	3
3	5
5	7
8	10

Step 1: Calculate the Means Calculate the mean of each feature:

$$\mu_X = \frac{2+3+5+8}{4} = 4.5, \quad \mu_Y = \frac{3+5+7+10}{4} = 6.25$$

Example

Consider a dataset with two features, X and Y:

X	Y
2	3
3	5
5	7
8	10

Step 1: Calculate the Means Calculate the mean of each feature:

$$\mu_X = \frac{2+3+5+8}{4} = 4.5, \quad \mu_Y = \frac{3+5+7+10}{4} = 6.25$$

Step 2: Center the Data Center the data by subtracting the mean:

$$X_{centered} = \begin{bmatrix} 2-4.5 & 3-6.25 \\ 3-4.5 & 5-6.25 \\ 5-4.5 & 7-6.25 \\ 8-4.5 & 10-6.25 \end{bmatrix} = \begin{bmatrix} -2.5 & -3.25 \\ -1.5 & -1.25 \\ 0.5 & 0.75 \\ 3.5 & 3.75 \end{bmatrix}$$

Now calculate the covariance matrix:

$$C = \frac{1}{N-1}(X_{centered})^T X_{centered}$$

Now calculate the covariance matrix:

$$C = \frac{1}{N-1}(X_{centered})^T X_{centered}$$

Calculating $(X_{centered})^T$:

$$(X_{centered})^T = \begin{bmatrix} -2.5 & -1.5 & 0.5 & 3.5 \\ -3.25 & -1.25 & 0.75 & 3.75 \end{bmatrix}$$

Now calculate the covariance matrix:

$$C = \frac{1}{N-1}(X_{centered})^T X_{centered}$$

Calculating $(X_{centered})^T$:

$$(X_{centered})^T = \begin{bmatrix} -2.5 & -1.5 & 0.5 & 3.5 \\ -3.25 & -1.25 & 0.75 & 3.75 \end{bmatrix}$$

Now compute the product:

$$(X_{centered})^T X_{centered} = \begin{bmatrix} (-2.5)(-2.5) + (-1.5)(-1.5) + (0.5)(0.5) + (3.5)(3.5) & (-2.5)(-3.25) + (-1.5)(-1.25) + (0.5)(0.75) + (3.5)(3.75) \\ (-3.25)(-2.5) + (-1.25)(-1.5) + (0.75)(0.5) + (3.75)(3.5) & (-3.25)(-3.25) + (-1.25)(-1.25) + (0.75)(0.75) + (3.75)(3.75) \end{bmatrix}$$

Now calculate the covariance matrix:

$$C = \frac{1}{N-1}(X_{centered})^T X_{centered}$$

Calculating $(X_{centered})^T$:

$$(X_{centered})^T = \begin{bmatrix} -2.5 & -1.5 & 0.5 & 3.5 \\ -3.25 & -1.25 & 0.75 & 3.75 \end{bmatrix}$$

Now compute the product:

$$(X_{centered})^T X_{centered} = \begin{bmatrix} (-2.5)(-2.5) + (-1.5)(-1.5) + (0.5)(0.5) + (3.5)(3.5) & (-2.5)(-3.25) + (-1.5)(-1.25) + (0.5)(0.75) + (3.5)(3.75) \\ (-3.25)(-2.5) + (-1.25)(-1.5) + (0.75)(0.5) + (3.75)(3.5) & (-3.25)(-3.25) + (-1.25)(-1.25) + (0.75)(0.75) + (3.75)(3.75) \end{bmatrix}$$

Calculating each term:

$$= \begin{bmatrix} 19.5 & 21.625 \\ 21.625 & 24.6875 \end{bmatrix}$$

Now calculate the covariance matrix:

$$C = \frac{1}{N-1}(X_{centered})^T X_{centered}$$

Calculating $(X_{centered})^T$:

$$(X_{centered})^T = \begin{bmatrix} -2.5 & -1.5 & 0.5 & 3.5 \\ -3.25 & -1.25 & 0.75 & 3.75 \end{bmatrix}$$

Now compute the product:

$$(X_{centered})^T X_{centered} = \begin{bmatrix} (-2.5)(-2.5) + (-1.5)(-1.5) + (0.5)(0.5) + (3.5)(3.5) & (-2.5)(-3.25) + (-1.5)(-1.25) + (0.5)(0.75) + (3.5)(3.75) \\ (-3.25)(-2.5) + (-1.25)(-1.5) + (0.75)(0.5) + (3.75)(3.5) & (-3.25)(-3.25) + (-1.25)(-1.25) + (0.75)(0.75) + (3.75)(3.75) \end{bmatrix}$$

Calculating each term:

$$= \begin{bmatrix} 19.5 & 21.625 \\ 21.625 & 24.6875 \end{bmatrix}$$

Finally, divide by $N-1=3$:

$$C = \frac{1}{3}\begin{bmatrix} 19.5 & 21.625 \\ 21.625 & 24.6875 \end{bmatrix} \approx \begin{bmatrix} 6.5 & 7.2083 \\ 7.2083 & 8.2292 \end{bmatrix}$$

Interpretation

- The diagonal elements C_{11} and C_{22} represent the variances of X and Y, respectively.

Interpretation

- The diagonal elements C_{11} and C_{22} represent the variances of X and Y, respectively.
- The off-diagonal elements C_{12} and C_{21} represent the covariance between X and Y.

Conclusion

- Variance and covariance are essential for understanding the relationships between variables.

Conclusion

- Variance and covariance are essential for understanding the relationships between variables.
- Centralizing data helps in simplifying analyses and calculations.

Conclusion

- Variance and covariance are essential for understanding the relationships between variables.
- Centralizing data helps in simplifying analyses and calculations.
- These concepts form the foundation for more advanced techniques like PCA.

Introduction

- Principal Component Analysis (PCA) is a dimensionality reduction technique.

Introduction

- Principal Component Analysis (PCA) is a dimensionality reduction technique.
- It transforms the data to a new coordinate system with axes (principal components) that maximize variance.

Introduction

- Principal Component Analysis (PCA) is a dimensionality reduction technique.
- It transforms the data to a new coordinate system with axes (principal components) that maximize variance.
- PCA is widely used in data preprocessing, visualization, and noise reduction.

Proof (Outline of PCA)

Centralized Data Matrix: Let D be the centralized data matrix:

$$D = X - \mathbf{1}\mu^T$$

Proof (Outline of PCA)

Centralized Data Matrix: Let D be the centralized data matrix:

$$D = X - \mathbf{1}\mu^T$$

Projection onto an Arbitrary Vector: Let $\mathbf{u}$ be an arbitrary unit vector. The projection is:

$$D_{\text{proj}} = D\mathbf{u}$$

Variance of the Projection: The variance can be expressed as:

$$\mathrm{Var}(D_{\mathrm{proj}}) = \frac{1}{N} \sum_{i=1}^{N} (D_i \cdot \mathbf{u})^2$$

Variance of the Projection: The variance can be expressed as:

$$\mathrm{Var}(D_{\mathrm{proj}}) = \frac{1}{N}\sum_{i=1}^{N}(D_i \cdot \mathbf{u})^2$$

Expressing Variance in Matrix Form:

$$\mathrm{Var}(D_{\mathrm{proj}}) = \frac{1}{N}(D\mathbf{u})^T(D\mathbf{u}) = \frac{1}{N}\mathbf{u}^T D^T D\mathbf{u}$$

Variance of the Projection: The variance can be expressed as:

$$\mathrm{Var}(D_{\text{proj}}) = \frac{1}{N}\sum_{i=1}^{N}(D_i \cdot \mathbf{u})^2$$

Expressing Variance in Matrix Form:

$$\mathrm{Var}(D_{\text{proj}}) = \frac{1}{N}(D\mathbf{u})^T(D\mathbf{u}) = \frac{1}{N}\mathbf{u}^T D^T D\mathbf{u}$$

Maximization of Variance: We want to maximize:

$$\max_{\mathbf{u}} \mathbf{u}^T S\mathbf{u}, \quad S = \frac{1}{N}D^T D$$

Rayleigh Quotient: The problem becomes finding the maximum of:

$$R(\mathbf{u}) = \frac{\mathbf{u}^T S \mathbf{u}}{\mathbf{u}^T \mathbf{u}} \quad \text{(unit vector)}$$

Rayleigh Quotient: The problem becomes finding the maximum of:

$$R(\mathbf{u}) = \frac{\mathbf{u}^T S \mathbf{u}}{\mathbf{u}^T \mathbf{u}} \quad \text{(unit vector)}$$

Eigenvalue Problem: The solution yields the eigenvalues and eigenvectors of S. The maximum occurs when $\mathbf{u}$ is the eigenvector corresponding to the largest eigenvalue.

Rayleigh Quotient: The problem becomes finding the maximum of:

$$R(\mathbf{u}) = \frac{\mathbf{u}^T S \mathbf{u}}{\mathbf{u}^T \mathbf{u}} \quad \text{(unit vector)}$$

Eigenvalue Problem: The solution yields the eigenvalues and eigenvectors of S. The maximum occurs when $\mathbf{u}$ is the eigenvector corresponding to the largest eigenvalue. **Conclusion:** Projecting onto the eigenvector corresponding to the largest eigenvalue achieves maximum variance, defining the principal component.

Example

Consider the following dataset with 5 observations and 2 features:

$$X = \begin{bmatrix} 2 & 3 \\ 3 & 4 \\ 4 & 5 \\ 5 & 6 \\ 6 & 7 \end{bmatrix}$$

Example

Consider the following dataset with 5 observations and 2 features:

$$X = \begin{bmatrix} 2 & 3 \\ 3 & 4 \\ 4 & 5 \\ 5 & 6 \\ 6 & 7 \end{bmatrix}$$

Step 1: Centering the Data First, we compute the mean of each feature:

$$\mu_1 = \frac{1}{5}(2+3+4+5+6) = 4$$

$$\mu_2 = \frac{1}{5}(3+4+5+6+7) = 5$$

Next, we center the data by subtracting the mean from each feature:

$$\tilde{X} = X - \begin{bmatrix} 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

Step 2: Covariance Matrix Now, we calculate the covariance matrix C:

Step 2: Covariance Matrix Now, we calculate the covariance matrix C:

$$C = \frac{1}{N-1}\tilde{X}^T\tilde{X} = \frac{1}{4}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

Step 2: Covariance Matrix Now, we calculate the covariance matrix C:

$$C = \frac{1}{N-1}\tilde{X}^T\tilde{X} = \frac{1}{4}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

Calculating $\tilde{X}^T\tilde{X}$:

$$\tilde{X}^T\tilde{X} = \begin{bmatrix} (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 & (-2)(-2) + (-1)(-1) + 0 + 1 + 2 \\ (-2)(-2) + (-1)(-1) + 0 + 1 + 2 & (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 \end{bmatrix}$$

Step 2: Covariance Matrix Now, we calculate the covariance matrix C:

$$C = \frac{1}{N-1}\tilde{X}^T\tilde{X} = \frac{1}{4}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

Calculating $\tilde{X}^T\tilde{X}$:

$$\tilde{X}^T\tilde{X} = \begin{bmatrix} (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 & (-2)(-2) + (-1)(-1) + 0 + 1 + 2 \\ (-2)(-2) + (-1)(-1) + 0 + 1 + 2 & (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 \end{bmatrix}$$

Thus, the covariance matrix C is:

$$C = \frac{1}{4}\begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix} = \begin{bmatrix} 2.5 & 2.5 \\ 2.5 & 2.5 \end{bmatrix}$$

Step 3: Eigenvalue Decomposition To find the eigenvalues and eigenvectors of C, we solve the characteristic equation:

Step 3: Eigenvalue Decomposition To find the eigenvalues and eigenvectors of C, we solve the characteristic equation:

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

where I is the identity matrix.

Step 3: Eigenvalue Decomposition To find the eigenvalues and eigenvectors of C, we solve the characteristic equation:

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

where I is the identity matrix.

$$\det\left(\begin{bmatrix} 2.5 - \lambda & 2.5 \\ 2.5 & 2.5 - \lambda \end{bmatrix}\right) = (2.5 - \lambda)(2.5 - \lambda) - (2.5)(2.5) = 0$$

Step 3: Eigenvalue Decomposition To find the eigenvalues and eigenvectors of C, we solve the characteristic equation:

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

where I is the identity matrix.

$$\det\left(\begin{bmatrix} 2.5 - \lambda & 2.5 \\ 2.5 & 2.5 - \lambda \end{bmatrix}\right) = (2.5 - \lambda)(2.5 - \lambda) - (2.5)(2.5) = 0$$

Expanding the determinant:

$$(2.5 - \lambda)^2 - 6.25 = 0 \implies \lambda^2 - 5\lambda = 0$$

Step 3: Eigenvalue Decomposition To find the eigenvalues and eigenvectors of C, we solve the characteristic equation:

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

where I is the identity matrix.

$$\det\left(\begin{bmatrix} 2.5 - \lambda & 2.5 \\ 2.5 & 2.5 - \lambda \end{bmatrix}\right) = (2.5 - \lambda)(2.5 - \lambda) - (2.5)(2.5) = 0$$

Expanding the determinant:

$$(2.5 - \lambda)^2 - 6.25 = 0 \implies \lambda^2 - 5\lambda = 0$$

Thus, the eigenvalues are:

$$\lambda_1 = 5, \quad \lambda_2 = 0$$

Next, we find the eigenvectors for each eigenvalue.

Next, we find the eigenvectors for each eigenvalue. For $\lambda_1 = 5$:

$$(C - 5I)\mathbf{v} = 0 \implies \begin{bmatrix} -2.5 & 2.5 \\ 2.5 & -2.5 \end{bmatrix}$$

Next, we find the eigenvectors for each eigenvalue. For $\lambda_1 = 5$:

$$(C - 5I)\mathbf{v} = 0 \implies \begin{bmatrix} -2.5 & 2.5 \\ 2.5 & -2.5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

This gives the eigenvector:

$$\mathbf{v_1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Next, we find the eigenvectors for each eigenvalue. For $\lambda_1 = 5$:

$$(C - 5I)\mathbf{v} = 0 \implies \begin{bmatrix} -2.5 & 2.5 \\ 2.5 & -2.5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

This gives the eigenvector:

$$\mathbf{v_1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For $\lambda_2 = 0$:

$$(C)\mathbf{v} = 0 \implies \begin{bmatrix} 2.5 & 2.5 \\ 2.5 & 2.5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

This gives the eigenvector:

$$\mathbf{v_2} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

We normalize the eigenvectors to have unit length:

We normalize the eigenvectors to have unit length: For $\mathbf{v_1}$:

$$\mathbf{v_1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

We normalize the eigenvectors to have unit length: For $\mathbf{v_1}$:

$$\mathbf{v_1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

For $\mathbf{v_2}$:

$$\mathbf{v_2} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Projection Matrix

$$P = v_1 v_1^T + v_2 v_2^T = \frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}$$

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

Step 5: Projection onto Principal Components Now we can project the centered data onto the principal components defined by $\mathbf{v_1}$ and $\mathbf{v_2}$:

$$Z_1 = \tilde{X}\mathbf{v_1} = \begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} \\ -1\sqrt{2} \\ 0 \\ 1\sqrt{2} \\ 2\sqrt{2} \end{bmatrix} = \begin{bmatrix} -2.83 \\ -1.41 \\ 0 \\ 1.41 \\ 2.83 \end{bmatrix}$$

$$Z_2 = \tilde{X}\mathbf{v_2} = \begin{bmatrix} -2 & -2 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

In this example, we calculated two principal components for the dataset. The first principal component captures the maximum variance, while the second component, in this case, turns out to be a direction with no variance (as the data is perfectly linear). This example illustrates how PCA helps in reducing dimensionality while retaining the structure of the data.

Python Implementation

```python
#Step 1: Import Libraries
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
#Step 2: Create Dataset
# Create a sample dataset
X = np.array([[2, 3],
              [3, 4],
              [4, 5],
              [5, 6],
              [6, 7]])
#Step 3: Center the Data
# Centering the data
X_centered = X - np.mean(X, axis=0)
#Step 4: Compute Covariance Matrix
# Compute the covariance matrix
cov_matrix = np.cov(X_centered, rowvar=False)
```

```
# Eigenvalue decomposition
eigenvalues, eigenvectors=np.linalg.eig(cov_matrix)
# Sort eigenvalues and eigenvectors
sorted_indices = np.argsort(eigenvalues)[::-1]
sorted_eigenvalues = eigenvalues[sorted_indices]
sorted_eigenvectors=eigenvectors[:, sorted_indices]
#Select the top \(k\) eigenvectors and project the
# Select the first two eigenvectors
k = 2
top_eigenvectors = sorted_eigenvectors[:, :k]
# Project the data
X_pca = X_centered.dot(top_eigenvectors)

# Visualize the original data and PCA result
plt.figure(figsize=(8, 6))
plt.scatter(X[:, 0], X[:, 1], color='blue', label='
plt.scatter(X_pca[:, 0], X_pca[:, 1], color='red',
plt.title('PCA Projection')
plt.xlabel('Principal Component 1')
```

```
plt.ylabel('Principal Component 2')
plt.legend()
plt.grid()
plt.show()
```

```
# Using scikit-learn for PCA
pca = PCA(n_components=2)
X_pca_sklearn = pca.fit_transform(X)

# Visualize the result
plt.figure(figsize=(8, 6))
plt.scatter(X[:, 0], X[:, 1], color='blue', label='
plt.scatter(X_pca_sklearn[:, 0], X_pca_sklearn[:, 1
plt.title('PCA using Scikit-Learn')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.grid()
plt.show()
```

Conclusion

- PCA reduces dimensionality while retaining most of the variance.
- This can be done manually or with libraries like scikit-learn.
- PCA is useful for visualization, data compression, and noise reduction.

The principal component captures the maximum variance in the dataset, reducing the dimensionality while retaining the structure of the data. In this example, we calculated the covariance matrix, performed eigenvalue decomposition, and projected the data onto the principal component.