PCA and Correlation in Data Analysis

Dr. Shailesh Sivan

June 10, 2025

What is PCA?

- PCA (Principal Component Analysis) is a statistical technique used for dimensionality reduction.
- It transforms the data to a new coordinate system:
  - Axes = directions of maximum variance (principal components).
  - First few PCs capture most information.
- Commonly used in preprocessing for ML models and visualizations.

Steps in PCA

1. Center the data (subtract the mean).
2. Compute the covariance matrix.
3. Compute eigenvalues and eigenvectors of the covariance matrix.
4. Sort eigenvectors by decreasing eigenvalues.
5. Select top k eigenvectors for dimensionality reduction.
6. Project data onto new basis.

Mathematical Formulation

- Let $X \in \mathbb{R}^{m \times n}$ be a centered data matrix.
- Covariance matrix:
$$\Sigma = \frac{1}{m} X^T X$$
- Eigen decomposition:
$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i$$
- Principal components are the eigenvectors $\mathbf{v}_i$.
- Projected data:
$$Z = XV_k$$

Numerical Example of PCA

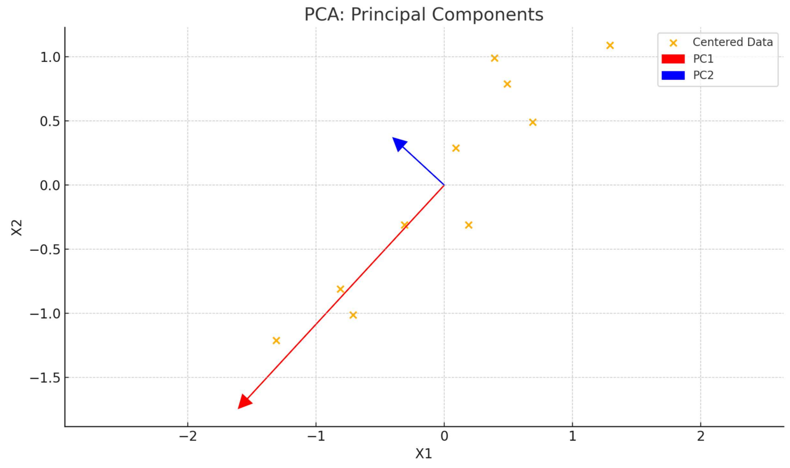
- Dataset: 2D data points (10 samples).
- Step 1: Center data by subtracting mean.
- Step 2: Compute covariance matrix.
- Step 3: Find eigenvalues & eigenvectors.
- Step 4: Project data onto top 1 PC (1D).

Projected Point

$$Z = X_{\text{centered}} \cdot \mathbf{v}_1$$

PCA Visualization

- The figure below shows the centered data and two principal directions.



Sample Dataset (2D)

Sample	x_1	x_2
A	2.5	2.4
B	0.5	0.7
C	2.2	2.9
D	1.9	2.2
E	3.1	3.0
F	2.3	2.7
G	2.0	1.6
H	1.0	1.1
I	1.5	1.6
J	1.1	0.9

Step 1: Mean Centering

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 1.81 \\ 1.91 \end{bmatrix}$$

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

Each value in the dataset is adjusted:

$$x_{ij}^{\text{centered}} = x_{ij} - \mu_j$$

Step 2: Covariance Matrix

$$\Sigma = \frac{1}{n-1} X^T X = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

Covariance matrix represents feature variances and their correlations.

Step 3: Eigen Decomposition

- Eigenvalues:

$$\lambda_1 = 1.2840, \quad \lambda_2 = 0.0490$$

- Corresponding Eigenvectors:

$$\mathbf{v}_1 = \begin{bmatrix} 0.6779 \\ 0.7352 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -0.7352 \\ 0.6779 \end{bmatrix}$$

Principal Component: Direction of maximum variance.

Step 4: Project Data onto PC1

- Project each centered point onto $\mathbf{v}_1$:

$$z_i = \mathbf{x}_i^{\text{centered}} \cdot \mathbf{v}_1$$

- Projected 1D values:

$$Z = \begin{bmatrix} 0.82797 \\ -1.77758 \\ 0.9922 \\ 0.27421 \\ 1.6758 \\ 0.91295 \\ 0.0991 \\ -1.1446 \\ -0.43805 \\ -1.2238 \end{bmatrix}$$

Step 5: Variance Retained

- Total variance = sum of eigenvalues.
- Retained variance from PC1:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.2840}{1.2840 + 0.0490} \approx 0.963$$

- PCA with 1 component retains about **96.3%** of the original variance.

Summary

- PCA reduces dimensionality while preserving most of the variance.
- In this example:
  - Data reduced from 2D to 1D.
  - 96.3% variance retained.
- PCA is powerful for visualization, noise reduction, and ML preprocessing.

Applications of PCA

- Dimensionality reduction (e.g., reduce from 1000 to 50 features).
- Visualization of high-dimensional data.
- Noise filtering and compression.
- Speeding up ML algorithms.
- Removing correlated features.

Correlation Analysis

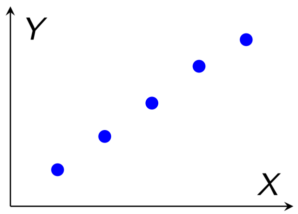
- Measures the strength and direction of linear relationship between two variables.
- Pearson Correlation Coefficient:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

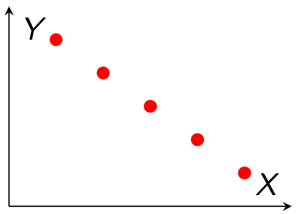
- $r \in [-1, 1]$: $+1$ perfect positive, -1 perfect negative, 0 no correlation

Positive and Negative Correlation

Positive Correlation



Negative Correlation



Direct Problem: Compute Correlation

- Let $X = [1, 2, 3, 4, 5]$, $Y = [2, 4, 5, 4, 5]$
- Mean: $\bar{X} = 3$, $\bar{Y} = 4$
- Numerator:

$$\sum (x_i - 3)(y_i - 4) = (1-3)(2-4) + \ldots = 6$$

- Denominator:

$$\sqrt{\sum (x_i - 3)^2} = \sqrt{10}, \quad \sqrt{\sum (y_i - 4)^2} = \sqrt{6}$$

- $r = \frac{6}{\sqrt{60}} \approx 0.77$

Zero Correlation Example

- Let $X = [1, 2, 3, 4, 5]$, $Y = [2, 2, 2, 2, 2]$
- $\mathrm{Var}(Y) = 0 \Rightarrow$ correlation is undefined or zero.
- No relationship can be detected using Pearson correlation.