# Principal Component Analysis (PCA)

Dr. Mrityunjoy Barman
**mrityunjoybarman@soa.ac.in**

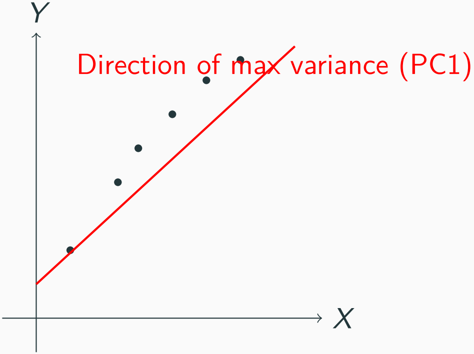
November 7, 2025

# Overview

- Motivation & intuition
- Z-score normalization
- Covariance matrix
- Eigenvalues & eigenvectors
- PCA algorithm (step-by-step)
- Worked numerical examples (2D, scaling effect, 3 variables)

# Why PCA? (Intuition)

- Reduce dimensionality while retaining maximal variance.
- Remove redundancy from correlated variables.
- Create uncorrelated axes (Principal Components, PCs).



# z-Score Normalization (Standardization)

We recall the standardization formula for the z-score normalization.

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

- Let $X = [x_1,\ x_2, \cdots, x_n]$ with mean $\bar{x}$ and standard deviation $s$.
- We transform each variable $X$ into $Z$ using (1).
- Ensures each variable contributes equally (mean $= 0$, SD $= 1$).
- PCA on the *correlation* matrix $\Leftrightarrow$ PCA on standardized data.

# Covariance Matrix

For $m$ numeric variables: we define the covariance matrix for the given set of variables $\{X_1,\ X_2, \cdots, X_m\}$ as $\Sigma = \left[\xi_{ij}\right]_{m\times m}$ with

$$\xi_{ij} = Cov(X_i,\ X_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{X}_i)(x_{kj} - \bar{X}_j)$$

and

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \cdots & \mathrm{Cov}(X_1, X_m) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m, X_1) & \cdots & \mathrm{Var}(X_m) \end{bmatrix} \tag{2}$$

# Eigenvalues and Eigenvectors

**Let $A$ be a square matrix. We call a number $\lambda$ to be an eigenvalue of $A$ if there is a vector $\mathbf{v} \neq 0$ such that $A\mathbf{v} = \lambda\mathbf{v}$.**

In case of PCA, the above notations have the following interpretations:

- $\mathbf{v}$: direction of a principal component (loading vector).
- $\lambda$: variance captured by that PC.
- **PCs are orthogonal; eigenvalues are nonnegative and sum to total variance**.
- $\sum_{k=1}^{m} \lambda_k = m =$ Total variance.

# PCA Algorithm (Step-by-Step)

1. Standardize the data variables $X_k \to Z_k$ (z-score normalization).
2. Compute covariance matrix (or correlation matrix) as

$$\Sigma = \begin{bmatrix} 1 & \cdots & \mathrm{Cov}(Z_1, Z_m) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(Z_m, Z_1) & \cdots & 1 \end{bmatrix} \tag{3}$$

   .
3. Compute the eigenvalues and their corresponding eigenvectors.
4. Sort eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ and shortlist them based on the preassigned criteria.
5. Compute PC scores: Taking linear combinations of the $Z$-variables with the help of the eigenvectors corresponding to the shortlisted eigenvalues.

# Worked Example 1: 2D Dataset (4 samples) i

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| 2 | 4 | 2 |
| 0 | 0 | 2 |
| 2 | 2 | 0 |
| 4 | 2 | 4 |

**Step 1:** Means: $\bar{X}_1 = 2,\ \bar{X}_2 = 2,\ \bar{X}_3 = 2$; and Std deviations: $\sigma_i = \sqrt{2}$, for $i = 1, 2, 3$.

# Worked Example 1: 2D Dataset (4 samples) ii

After standardization, the variables are transformed as

$$\begin{array}{c|c|c} Z_1 & Z_2 & Z_3 \\ \hline 0 & \sqrt{2} & 0 \\ -\sqrt{2} & -\sqrt{2} & 0 \\ 0 & 0 & -\sqrt{2} \\ \sqrt{2} & 0 & \sqrt{2} \end{array} \tag{4}$$

**Step 2:** The covariance (correlation) matrix:

$$\rho = \begin{bmatrix} 1 & 1/2 & -1/2 \\ 1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix} \tag{5}$$

# Worked Example 1: 2D Dataset (4 samples) iii

**Step 3:** The eigenvalues of $\rho$ are $\lambda_1 = 1 + \frac{1}{\sqrt{2}} \approx 1.7071,\ \lambda_2 = 1,\ \lambda_3 = 1 - \frac{1}{\sqrt{2}} \approx 0.2929$ in decreasing order. We shortlist only $\lambda_1$ and $\lambda_2$. The contribution of these eigenvalues to the total variance are given by

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ |
|---|---|---|
| $\frac{1.7071}{3}$ $= 56.90\%$ | $\frac{1}{3}$ $=33.33\%$ | $\frac{0.2929}{3}$ $=9.76\%$ |

# Worked Example 1: 2D Dataset (4 samples) iv

**Step 4:** The eigenvectors corresponding the shortlisted eigenvalues as follows:

$$e_1 = \begin{bmatrix} \sqrt{2} \\ 1 \\ -1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \tag{6}$$

**Step 5:** The principal components are now computed as (by taking linear combinations of $Z_1,\ Z_2,\ Z_3$)

$$Y_1 = \sqrt{2} Z_1 + Z_2 - Z_3, \quad Y_2 = Z_2 + Z_3. \tag{7}$$

# Summary

- Any Questions? Contact me at mrityunjoybarman@soa.ac.in.

# Principal Component Analysis (PCA)

Dr. Mrityunjoy Barman
**mrityunjoybarman@soa.ac.in**

November 7, 2025