Principal Component Analysis (PCA)

PCA

- Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in data analysis and machine learning.
- It transforms a dataset into a new coordinate system where the most significant variance is captured in the first few principal components.
- Principal Component Analysis (PCA) transforms a dataset into a new coordinate system where the axes (principal components) are aligned with the directions of **maximum variance** in the data.
- The first few principal components capture the most significant variations, while later components represent diminishing amounts of variance.

PCA

- **Variance** in data refers to the spread or dispersion of data points around the mean.
- In PCA, the goal is to **maximize variance** along new axes (principal components) so that the most important patterns in the data are preserved.
- Each **principal component (PC)** is a linear combination of the original features and captures a specific amount of variance.
- The **first principal component (PC1)** captures the **highest variance**, the **second principal component (PC2)** captures the second highest variance, and so on.

PCA

- Mathematically, if X is a centered dataset (zero mean), and C is its covariance matrix:

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

The **eigenvalues** $\lambda_1, \lambda_2, \ldots, \lambda_n$ of C represent the amount of variance captured by each principal component.

- The **first principal component (PC1)** has the highest eigenvalue λ_1, meaning it captures the most variance.
- The **second principal component (PC2)** has the second highest eigenvalue λ_2, and so on.
- The **total variance in the data** is:

$$\sum_{i=1}^{N} \lambda_i$$

Why does PCA Capture Most Variance in the First Few Components?

Maximizing Variance

- PCA finds new directions (principal components) such that the projected data retains maximum variance.
- Since variance corresponds to information, retaining variance means retaining key patterns in the data.

Decorrelation of Features

- PCA removes correlations between original features, making principal components **uncorrelated**.
- Since correlated variables contain redundant information, PCA merges them into fewer components.

Dimensionality Reduction Without Much Information Loss

- In high-dimensional data, many features may have small contributions to overall variance.
- The first few principal components capture most of the variance, allowing us to reduce dimensions while preserving essential information.

PCA - Steps

- Standardize the Data - Since PCA is affected by scale, it is common to standardize the dataset before applying it.
- Compute the Covariance Matrix of the dataset - to understand the relationships between features.
- Find Eigenvalues and Eigenvectors of the covariance matrix - The eigenvalues represent the variance captured by each principal component, while eigenvectors define their direction.
- Sort Eigenvalues in Descending Order and select the top k eigenvectors - By selecting the top k principal components (those with the highest eigenvalues), PCA reduces the number of features while retaining most of the dataset's variability.
- Project the Original Data onto the new feature space.

PCA

- Let's assume we have a dataset with m observations and n features, represented as an $m \times n$ matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

where each row represents an **observation**, and each column represents a **feature**.

PCA - Standardization (Mean Centering and Scaling)

- Since PCA is affected by scale, standardize the dataset by centering the features around the mean and normalizing their variance.
- For each feature j, compute:
  - The mean:

$$\mu_j = \frac{1}{M}\sum_{i=1}^{m} x_{ij}$$

  - The standard deviation:

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{n}\left(x_{ij} - \mu_j\right)^2}$$

- Now, transform each feature as:

$$X'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

- This ensures that each feature has zero mean and unit variance.
- This transformation ensures that all features contribute equally to the PCA analysis.

- Mean centering ensures that each feature has a mean of **zero**, which is important for PCA because it relies on the **covariance matrix**, which is computed around zero.
- scale each feature so that all variables have a **variance of 1**.
- Subtracting the mean from each value will center each feature around zero.

PCA - Compute the Covariance Matrix

- The covariance matrix captures the relationships between features:

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

where C is an n×n symmetric matrix. The element C_{ij} represents the covariance between the ith and jth features.

- For two features X_i and X_j, the covariance is:

$$C_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$$

If $C_{ij}>0$, the features are positively correlated.
If $C_{ij}<0$, the features are negatively correlated.
If $C_{ij}=0$, the features are uncorrelated.

PCA - Compute Eigenvalues and Eigen vectors

- PCA finds **principal components** by solving the eigenvalue problem:

$$Cv=\lambda v$$

where v is an **eigenvector** (direction of the principal component),

λ is the corresponding **eigenvalue** (amount of variance captured).

- To solve for v and λ, we solve: $\det(C-\lambda I)=0$ where I is the identity matrix.
- Each eigenvector represents a **principal component**, and its corresponding eigenvalue quantifies how much variance it captures.

PCA - Select the Top Principal Components

- Sort eigenvalues $\lambda1 \geq \lambda2 \geq \cdots \geq \lambda n$.
- Choose the top k eigenvectors corresponding to the largest k eigenvalues.
- The proportion of variance explained by the i-th principal component is:

$$\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$$

- We typically select the smallest k such that the cumulative variance exceeds a threshold (e.g., 95%).

PCA- Project Data onto the New Feature Space

- The new feature space (principal components) is:

$$Z = X'V_k$$

where:

- V_k is the matrix of the top k eigenvectors,
- Z is the transformed dataset with reduced dimensions.
- Each row of Z represents an observation in the reduced-dimensional space.

PCA – Solved Numerical

Consider a dataset with two features:

Observation	Feature 1(Age)	Feature 2 (Income)
1	25	50,000
2	30	60,000
3	35	70,000
4	40	80,000
5	45	90,000

PCA- Solved Numerical

Standardization

Compute Mean:

Mean of **Age**:

$$\mu_1 = \frac{25 + 30 + 35 + 40 + 45}{5} = 35$$

Mean of **Income**:

$$\mu_2 = \frac{50000 + 60000 + 70000 + 80000 + 90000}{5} = 70000$$

Compute Standard Deviation:

Standard deviation of **Age**:

$$\sigma_1 = \sqrt{\frac{(25-35)^2 + (30-35)^2 + (35-35)^2 + (40-35)^2 + (45-35)^2}{4}} = 7.91$$

Standard deviation of **Income**:

$$\sigma_2 = \sqrt{\frac{(50000-70000)^2 + (60000-70000)^2 + (70000-70000)^2 + (80000-70000)^2 + (90000-70000)^2}{4}} = 15811.39$$

PCA- Solved Numerical

Apply standardization

Standardized Age	Standardized Income
$\frac{25-35}{7.91} = -1.26$	$\frac{50000-70000}{15811.39} = -1.26$
$\frac{30-35}{7.91} = -0.63$	$\frac{60000-70000}{15811.39} = -0.63$
$\frac{35-35}{7.91} = 0.00$	$\frac{70000-70000}{15811.39} = 0.00$
$\frac{40-35}{7.91} = 0.63$	$\frac{80000-70000}{15811.39} = 0.63$
$\frac{45-35}{7.91} = 1.26$	$\frac{90000-70000}{15811.39} = 1.26$

The standardized dataset is:

-1.26	-1.26
-0.63	-0.63
0.00	0.00
0.63	0.63
1.26	1.26

PCA- Solved Numerical

Compute Covariance Matrix

- The covariance matrix measures the relationships between different features

$$C = \frac{1}{m-1} X_{\text{std}}^T X_{\text{std}}$$

- Computing Manually

$$C = \begin{bmatrix} 1.58 & 1.58 \\ 1.58 & 1.58 \end{bmatrix}$$

Since both features are highly correlated, there are strong covariance values.

PCA- Solved Numerical

- **Compute Eigenvalues and Eigenvectors**
- Eigenvalues tell us the amount of variance captured by each principal component, and eigenvectors define the directions.
- Solving **det(C−λI)=0**

$$\begin{vmatrix} 1.58-\lambda & 1.58 \\ 1.58 & 1.58-\lambda \end{vmatrix} = 0$$

Expanding determinant: $(1.58-\lambda)(1.58-\lambda)-(1.58)(1.58)=0$

$$\lambda^2 - 3.16\lambda + (2.5-2.5) = 0$$

$$\lambda^2 - 3.16\lambda = 0$$

$$\lambda(\lambda - 3.16) = 0$$

Thus, the eigenvalues are: λ1=3.16, λ2=0

PCA- Solved Numerical

Compute Eigenvectors

- For λ_1=3.16
- Solving (C−3.16I)v=0

$$\begin{bmatrix} 1.58 - 3.16 & 1.58 \\ 1.58 & 1.58 - 3.16 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -1.58 & 1.58 \\ 1.58 & -1.58 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

- Solving for v_1 and v_2, we get: v_1=v_2
- So the first principal component (normalized) is: $\mathbf{v_1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- For λ_2=0: $\mathbf{v_2} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

PCA- Solved Numerical

- **Transform the Data**
- Project the standardized data onto the principal components:

$$X_{pca} = X_{std} V \quad \text{where}$$

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Multiplying:

$$X_{pca} = \begin{bmatrix} -1.26 & -1.26 \\ -0.63 & -0.63 \\ 0.00 & 0.00 \\ 0.63 & 0.63 \\ 1.26 & 1.26 \end{bmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} -1.78 & 0 \\ -0.89 & 0 \\ 0.00 & 0 \\ 0.89 & 0 \\ 1.78 & 0 \end{bmatrix}$$

- Thus, the data is now one-dimensional along **PC1**, meaning PCA has effectively reduced the dimensionality.

PCA- Solved Numerical

Variance Explained

- The proportion of variance explained is:

$$\text{Explained Variance} = \frac{\lambda_i}{\sum \lambda}$$

PC1 Variance = 3.16 / (3.16+0)
=1.0

- **Standardization** ensured equal feature contribution.
- PCA computed new axes (PCs) that maximize variance.
- Only **PC1** is needed for dimensionality reduction.

Since **100% variance is captured in PC1**, we can discard PC2.

END