PRINCIPAL COMPONENT ANALYSIS (PCA)

- **Principal Component Analysis (PCA)** is a powerful dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving as much variance as possible.
- Used extensively in machine learning, image processing, and exploratory data analysis, PCA helps simplify complex datasets, improve model performance, and enhance visualization

Need for PCA

- **Reduces redundancy** → Correlated features carry overlapping information; PCA compresses them into fewer dimensions.
- **Mitigates overfitting** → By removing noise and irrelevant features, models generalize better.
- **Improves efficiency** → Lower dimensionality means faster training and reduced memory usage.
- **Enhances visualization** → Complex datasets can be projected into 2D or 3D for easier interpretation.

Why is High Dimensionality a Problem?

- **Computational inefficiency**: More features mean longer processing times and higher memory consumption.
- **Overfitting**: Models tend to learn noise rather than meaningful patterns.
- **Visualization challenges**: Understanding relationships in high-dimensional space is difficult.
- PCA addresses these issues by reducing dimensions while retaining the most important patterns in the data. It helps uncover latent structures, making datasets more manageable and interpretable.

How Does PCA Work?

- **Finding the Best Directions for Variance**
- PCA identifies the directions (or axes) along which the data varies the most.
- These directions are **orthogonal (perpendicular) to each other** and are ranked based on how much variance they capture.
- **Projecting Data onto Principal Components**
- The data is rotated and projected onto the new axes (principal components).
- The first principal component captures the highest variance, the second captures the next highest variance, and so on.
- **Dimensionality Reduction Without Losing Key Information**
- We keep only the top **k** principal components that retain most of the original variance.
- This reduces the number of dimensions while preserving the most essential patterns.

- **Mathematical Foundation of PCA**

Step 1: Standardizing the Data

- PCA is affected by differences in scale, so we first **standardize the dataset** by subtracting the mean and dividing by the standard deviation for each feature:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

- where:
- X is the original data matrix
- μ is the mean of each feature
- σ is the standard deviation of each feature

Step 2: Computing the Covariance Matrix

- The covariance matrix captures relationships between features. For a dataset with **n features**, the covariance matrix is given by:

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

where:

- C is an $n \times n$ matrix
- Each element C_{ij} represents the covariance between feature i and feature j.

- A high covariance means the two features are correlated, and PCA aims to **eliminate redundancy** by transforming data into uncorrelated principal components.

Step 3: Calculating Eigenvalues and Eigenvectors

- Eigenvalues and eigenvectors help determine the **principal components** of the data:
- **Eigenvectors** represent the new axes (directions of maximum variance).
- **Eigenvalues** measure how much variance each eigenvector captures.
- By solving the equation:
    - **Cv=λv**
- we obtain:
- **Eigenvalues (λ)**: Indicate the importance of each principal component.

- **Eigenvectors (v)**: Define the directions of new feature space.

Step 4: Selecting the Top k Principal Components

- To reduce dimensionality, we:
- Sort eigenvalues in **descending order** (highest variance first).
- Select the top **k** eigenvectors corresponding to the largest eigenvalues.
- Form a **projection matrix W** using these eigenvectors.
- W=[v1,v2,…,vk]
- where k<nk < nk<n, reducing the number of dimensions while retaining the most important variance.

Step 5: Transforming Data into New Feature Space

- Finally, we project the original dataset onto the selected principal components:

$$X_{\text{PCA}} = X_{\text{scaled}} W$$

- This transformed dataset, XPCA , now has reduced dimensions while preserving meaningful patterns.

- **5. Geometric Interpretation of PCA**
- Understanding PCA geometrically provides deeper insight into how it transforms data.
- PCA is essentially an **orthogonal projection** technique that finds new axes (principal components) to represent data in a lower-dimensional space while preserving its variance.

What are the advantages and disadvantages of PCA?

Advantages of PCA:

- **Easy to compute**. PCA is based on linear algebra, which is computationally easy to solve by computers.
- **Speeds up other machine learning algorithms**. Machine learning algorithms converge faster when trained on principal components instead of the original dataset.
- **Counteracts the issues of high-dimensional data**. High-dimensional data causes regression-based algorithms to overfit easily. By using PCA beforehand to lower the dimensions of the training dataset, we prevent the predictive algorithms from overfitting.

Disadvantages of PCA:

- **Low interpretability of principal components**. Principal components are linear combinations of the features from the original data, but they are not as easy to interpret. For example, it is difficult to tell which are the most important features in the dataset after computing principal components.

- **The trade-off between information loss and dimensionality reduction**. Although dimensionality reduction is useful, it comes at a cost. Information loss is a necessary part of PCA. Balancing the trade-off between dimensionality reduction and information loss is unfortunately a necessary compromise that we have to make when using PCA.

What are the assumptions and limitations of PCA?

- PCA is related to the set of operations in the Pearson correlation, so it inherits similar assumptions and limitations:
- **PCA assumes a correlation between features**. If the features (or dimensions or columns, in tabular data) are not correlated, PCA will be unable to determine principal components.
- **PCA is sensitive to the scale of the features**. Imagine we have two features - one takes values between 0 and 1000, while the other takes values between 0 and 1. PCA will be extremely biased towards the first feature being the first principle component, regardless of the *actual* maximum variance within the data. This is why it's so important to standardize the values first.
- **PCA is not robust against outliers**. Similar to the point above, the algorithm will be biased in datasets with strong outliers. This is why it is recommended to remove outliers before performing PCA.
- **PCA assumes a linear relationship between features**. The algorithm is not well suited to capturing non-linear relationships. That's why it's advised to turn non-linear features or relationships between features into linear, using the standard methods such as log transforms.
- **Technical implementations often assume no missing values**. When computing PCA using statistical software tools, they often assume that the feature set has no missing values (no empty rows). Be sure to remove those rows and/or columns with missing values, or impute missing values with a close approximation (e.g. the mean of the column).