Machine Learning 101

Rajdeep Chatterjee, Ph.D.
Amygdala AI, Bhubaneswar, India *

February 2025

Principal Component Analysis (PCA)

1 PCA Visualization

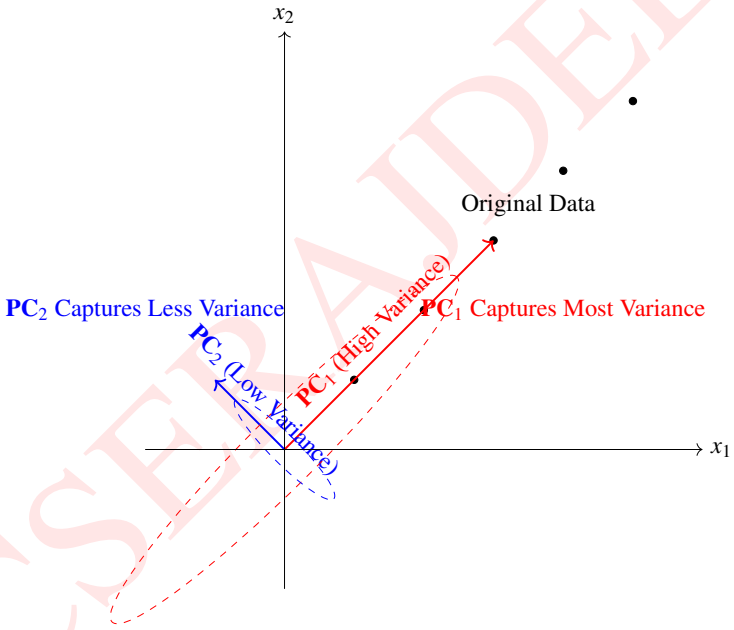

Figure 1: This visualization helps in understanding how PCA identifies the directions of maximum variance and reduces dimensionality while preserving the structure of the data.

2 The Blue Dashed Ellipse in PCA

In Principal Component Analysis (PCA), the **Blue dashed ellipse** represents the variance captured by the second principal component ($\mathbf{PC}_2$). Below is a detailed explanation of its meaning and significance.

*Amygdala AI, is an international volunteer-run research group that advocates for *AI for a better tomorrow* `http://amygdalaai.org/`.

1. What is the Blue Dashed Ellipse?

The Blue dashed ellipse is a geometric representation of the variance captured by the second principal component ($\mathbf{PC}_2$). In PCA, the variance of the data along each principal component is visualized using ellipses. The size and shape of the ellipse indicate how much variance is captured by that component.

- **Ellipse Orientation**: The ellipse is aligned with the direction of $\mathbf{PC}_2$.
- **Ellipse Size**: The size of the ellipse corresponds to the amount of variance captured by $\mathbf{PC}_2$.

2. Why is it Smaller?

The Blue dashed ellipse is much smaller than the red dashed ellipse (which represents the variance captured by $\mathbf{PC}_1$). This indicates that $\mathbf{PC}_2$ captures significantly less variance compared to $\mathbf{PC}_1$. Here's why:

- **PCA Orders Components by Variance**: PCA sorts the principal components in descending order of the variance they capture. The first principal component ($\mathbf{PC}_1$) captures the most variance, and each subsequent component captures less variance.
- **Low Variance in $\mathbf{PC}_2$**: In this example, $\mathbf{PC}_2$ captures the remaining variance that is orthogonal to $\mathbf{PC}_1$. If the data is highly correlated (e.g., lies close to a straight line in 2D), $\mathbf{PC}_2$ will capture very little variance, resulting in a small ellipse.

3. What Does This Mean in Practice?

The small size of the Blue dashed ellipse has important implications:

- **Dimensionality Reduction**: Since $\mathbf{PC}_2$ captures little variance, it can often be discarded without losing much information. This is the essence of dimensionality reduction in PCA.
- **Data Structure**: The small ellipse indicates that the data is primarily spread along $\mathbf{PC}_1$, meaning the data has a strong linear relationship or lies mostly along one direction.
- **Noise or Redundancy**: The variance captured by $\mathbf{PC}_2$ might represent noise or redundant information in the data.

4. Mathematical Interpretation

The size of the ellipse is determined by the **eigenvalue** (λ_2) associated with $\mathbf{PC}_2$. The eigenvalue represents the amount of variance captured by the component. If λ_2 is small, the ellipse will be small, indicating low variance.

$$\text{Variance captured by } \mathbf{PC}_2 = \lambda_2$$

5. Visualization Context

In the diagram:

- The **red dashed ellipse** (aligned with $\mathbf{PC}_1$) is large, indicating high variance.
- The **Blue dashed ellipse** (aligned with $\mathbf{PC}_2$) is small, indicating low variance.

The **Blue dashed ellipse** represents the variance captured by the second principal component ($\mathbf{PC}_2$). Its small size indicates that $\mathbf{PC}_2$ captures much less variance compared to $\mathbf{PC}_1$, which is a key insight for dimensionality reduction and understanding the structure of the data.

3 PCA on a Dummy Dataset

Step 1: Define the Dataset-1

Consider the following dataset with 2 features (columns) and 4 samples (rows):

$$\mathbf{X} = \begin{bmatrix} 4 & 11 \\ 8 & 4 \\ 13 & 5 \\ 7 & 14 \end{bmatrix}$$

Step 2: Center the Data

PCA requires the data to be centered around the mean. Compute the mean of each column and subtract it from the data.

$$\mu = \begin{bmatrix} 8.0 & 8.5 \end{bmatrix}$$

$$\mathbf{X}_{\text{centered}} = \mathbf{X} - \mu = \begin{bmatrix} 4-8.0 & 11-8.5 \\ 8-8.0 & 4-8.5 \\ 13-8.0 & 5-8.5 \\ 7-8.0 & 14-8.5 \end{bmatrix} = \begin{bmatrix} -4.0 & 2.5 \\ 0.0 & -4.5 \\ 5.0 & -3.5 \\ -1.0 & 5.5 \end{bmatrix}$$

Step 3: Compute the Covariance Matrix

The covariance matrix $\mathbf{C}$ is given by:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_{\text{centered}}^T \mathbf{X}_{\text{centered}}$$

$$\mathbf{C} = \frac{1}{3}\begin{bmatrix} -4.0 & 0.0 & 5.0 & -1.0 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix} \begin{bmatrix} -4.0 & 2.5 \\ 0.0 & -4.5 \\ 5.0 & -3.5 \\ -1.0 & 5.5 \end{bmatrix}$$

$$\mathbf{C} = \frac{1}{3}\begin{bmatrix} 42.0 & -26.0 \\ -26.0 & 64.5 \end{bmatrix} = \begin{bmatrix} 14.0 & -8.6667 \\ -8.6667 & 21.5 \end{bmatrix}$$

Step 4: Compute Eigenvalues and Eigenvectors

Find the eigenvalues λ and eigenvectors $\mathbf{v}$ of the covariance matrix $\mathbf{C}$.

$$\det(\mathbf{C} - \lambda \mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} 14.0-\lambda & -8.6667 \\ -8.6667 & 21.5-\lambda \end{bmatrix}\right) = 0$$

$$(14.0-\lambda)(21.5-\lambda) - (-8.6667)^2 = 0$$

$$\lambda^2 - 35.5\lambda + 301.0 - 75.1111 = 0$$

$$\lambda^2 - 35.5\lambda + 225.8889 = 0$$

Solving the quadratic equation:

$$\lambda = \frac{35.5 \pm \sqrt{(35.5)^2 - 4 \cdot 1 \cdot 225.8889}}{2}$$

$$\lambda = \frac{35.5 \pm \sqrt{1260.25 - 903.5556}}{2}$$

$$\lambda = \frac{35.5 \pm \sqrt{356.6944}}{2}$$

$$\lambda = \frac{35.5 \pm 18.886}{2}$$

The eigenvalues are:

$$\lambda_1 = \frac{35.5 + 18.886}{2} = 27.193, \quad \lambda_2 = \frac{35.5 - 18.886}{2} = 8.307$$

The corresponding eigenvectors are:

$$\mathbf{v}_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 5: Sort Eigenvalues and Eigenvectors

Sort the eigenvalues in descending order and arrange the corresponding eigenvectors accordingly.

$$\lambda_1 = 27.193, \quad \lambda_2 = 8.307$$

$$\mathbf{v}_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 6: Project the Data onto the Principal Components

The principal components are the eigenvectors corresponding to the largest eigenvalues. Project the centered data onto the principal components.

$$\mathbf{P} = \mathbf{X}_{\text{centered}} \cdot \mathbf{v}_1$$

$$\mathbf{P} = \begin{bmatrix} -4.0 & 2.5 \\ 0.0 & -4.5 \\ 5.0 & -3.5 \\ -1.0 & 5.5 \end{bmatrix} \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix} = \begin{bmatrix} -4.0 \cdot 0.5574 + 2.5 \cdot (-0.8303) \\ 0.0 \cdot 0.5574 + (-4.5) \cdot (-0.8303) \\ 5.0 \cdot 0.5574 + (-3.5) \cdot (-0.8303) \\ -1.0 \cdot 0.5574 + 5.5 \cdot (-0.8303) \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} -2.2296 - 2.07575 \\ 0.0 + 3.73635 \\ 2.787 + 2.90605 \\ -0.5574 - 4.56665 \end{bmatrix} = \begin{bmatrix} -4.30535 \\ 3.73635 \\ 5.69305 \\ -5.12405 \end{bmatrix}$$

Step 7: Final Transformed Data

The transformed data in the principal component space is:

$$\mathbf{X}_{\text{PCA}} = \begin{bmatrix} -4.30535 \\ 3.73635 \\ 5.69305 \\ -5.12405 \end{bmatrix}$$

Second Example

Step 1: Define the Dataset-2

Consider the following dummy dataset with 2 features (columns) and 4 samples (rows):

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \\ 4 & 5 \end{bmatrix}$$

Step 2: Center the Data

PCA requires the data to be centered around the mean. Compute the mean of each column and subtract it from the data.

$$\mu = \begin{bmatrix} 2.5 & 3.5 \end{bmatrix}$$

$$\mathbf{X}_{\text{centered}} = \mathbf{X} - \mu = \begin{bmatrix} 1-2.5 & 2-3.5 \\ 2-2.5 & 3-3.5 \\ 3-2.5 & 4-3.5 \\ 4-2.5 & 5-3.5 \end{bmatrix} = \begin{bmatrix} -1.5 & -1.5 \\ -0.5 & -0.5 \\ 0.5 & 0.5 \\ 1.5 & 1.5 \end{bmatrix}$$

Step 3: Compute the Covariance Matrix

The covariance matrix $\mathbf{C}$ is given by:

$$\mathbf{C} = \frac{1}{n-1} \mathbf{X}_{\text{centered}}^T \mathbf{X}_{\text{centered}}$$

$$\mathbf{C} = \frac{1}{3} \begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -1.5 & -0.5 & 0.5 & 1.5 \end{bmatrix} \begin{bmatrix} -1.5 & -1.5 \\ -0.5 & -0.5 \\ 0.5 & 0.5 \\ 1.5 & 1.5 \end{bmatrix}$$

$$\mathbf{C} = \frac{1}{3}\begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} = \begin{bmatrix} 1.6667 & 1.6667 \\ 1.6667 & 1.6667 \end{bmatrix}$$

Step 4: Compute Eigenvalues and Eigenvectors

Find the eigenvalues λ and eigenvectors $\mathbf{v}$ of the covariance matrix $\mathbf{C}$.

$$\det(\mathbf{C} - \lambda \mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} 1.6667 - \lambda & 1.6667 \\ 1.6667 & 1.6667 - \lambda \end{bmatrix}\right) = 0$$

$$(1.6667 - \lambda)^2 - (1.6667)^2 = 0$$

$$\lambda^2 - 3.3334\lambda = 0 \implies \lambda(\lambda - 3.3334) = 0$$

The eigenvalues are:

$$\lambda_1 = 3.3334, \quad \lambda_2 = 0$$

The corresponding eigenvectors are:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Step 5: Sort Eigenvalues and Eigenvectors

Sort the eigenvalues in descending order and arrange the corresponding eigenvectors accordingly.

$$\lambda_1 = 3.3334, \quad \lambda_2 = 0$$

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Step 6: Project the Data onto the Principal Components

The principal components are the eigenvectors corresponding to the largest eigenvalues. Project the centered data onto the principal components.

$$\mathbf{P} = \mathbf{X}_{\text{centered}} \cdot \mathbf{v}_1$$

$$\mathbf{P} = \begin{bmatrix} -1.5 & -1.5 \\ -0.5 & -0.5 \\ 0.5 & 0.5 \\ 1.5 & 1.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix}$$

Step 7: Final Transformed Data

The transformed data in the principal component space is:

$$\mathbf{X}_{\text{PCA}} = \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix}$$

4 Zero-Mean Dataset

A **zero-mean dataset** is a dataset where the mean of each feature (dimension) is zero. This is achieved by *centering* the data, which involves subtracting the mean of each feature from all data points.

4.1 Importance of Zero-Mean Data in PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the directions (principal components) along which variance in the data is maximized. The reasons why zero-mean data is crucial in PCA are:

4.2 Ensures Correct Covariance Computation

PCA relies on the **covariance matrix** to determine the directions of maximum variance. The covariance between two features X_i and X_j is computed as:

$$\text{Cov}(X_i, X_j) = \frac{1}{n} \sum_{k=1}^{n} (X_{ik} - \mu_i)(X_{jk} - \mu_j) \tag{1}$$

If the data is not centered (i.e., $\mu_i \neq 0$), the covariance computation will be incorrect, leading to incorrect principal components.

4.3 Aligns PCA with Eigenvalue Decomposition

PCA finds principal components by performing **eigenvalue decomposition (EVD)** or **singular value decomposition (SVD)** on the covariance matrix. If the dataset is not centered, the eigenvectors may not correctly capture variance directions.

4.4 Improves Numerical Stability

A dataset with large mean values may cause numerical instability in matrix computations, making PCA less reliable.

4.5 Zero-Centering Data

To center a dataset X with m features:

$$X_{\text{centered}} = X - \mu \tag{2}$$

where μ is the mean of each feature column.

5 Merits and Demerits of PCA

Merits of PCA

- **Dimensionality Reduction**: PCA reduces the number of features (dimensions) in a dataset while retaining most of the variance in the data. This simplifies the dataset and makes it easier to visualize and analyze.
- **Noise Reduction**: By focusing on the principal components (directions of maximum variance), PCA can filter out noise and irrelevant features, improving the quality of the data.
- **Improves Computational Efficiency**: Reducing the number of features decreases the computational cost of training machine learning models, especially for large datasets.
- **Uncorrelated Features**: PCA transforms the original features into a set of uncorrelated principal components, which can be beneficial for algorithms that assume independence between features.
- **Data Visualization**: PCA can reduce high-dimensional data to 2 or 3 dimensions, making it easier to visualize and interpret.
- **Feature Extraction**: PCA can be used as a feature extraction technique to identify the most important features in a dataset.

Demerits of PCA

- **Loss of Interpretability**: The principal components are linear combinations of the original features, which can make them difficult to interpret in terms of the original variables.
- **Linear Assumption**: PCA assumes that the relationships between variables are linear. It may not perform well on datasets with non-linear relationships.
- **Sensitive to Scaling**: PCA is sensitive to the scale of the input features. Features with larger scales can dominate the principal components, so data must be standardized before applying PCA.
- **Information Loss**: While PCA retains most of the variance, some information is inevitably lost when reducing dimensions. This can negatively impact the performance of downstream tasks.
- **Outliers**: PCA is sensitive to outliers, as they can significantly affect the directions of maximum variance and distort the principal components.
- **Not Suitable for All Datasets**: PCA may not be effective for datasets with categorical variables or datasets where the variance is not a good measure of importance.

6 Benefits of Orthogonality in PCA

Principal Component Analysis (PCA) relies on the orthogonality of its principal components (eigenvectors) to achieve dimensionality reduction effectively. The orthogonality property provides several key benefits:

1. Uncorrelated Features

The principal components are orthogonal to each other, meaning they are uncorrelated. Mathematically, for two principal components $\mathbf{PC}_i$ and $\mathbf{PC}_j$:

$$\mathbf{PC}_i \cdot \mathbf{PC}_j = 0 \quad \text{for} \quad i \neq j.$$

This ensures that each principal component captures unique information, eliminating redundancy in the data.

2. Maximization of Variance

Orthogonality ensures that the variance captured by each principal component is maximized in a sequential manner. The first principal component ($\mathbf{PC}_1$) captures the maximum variance, the second principal component ($\mathbf{PC}_2$) captures the maximum remaining variance orthogonal to $\mathbf{PC}_1$, and so on. This can be expressed as:

$$\text{Var}(\mathbf{PC}_1) \geq \text{Var}(\mathbf{PC}_2) \geq \cdots \geq \text{Var}(\mathbf{PC}_k).$$

3. Simplification of Projection

Orthogonal principal components form an orthonormal basis for the transformed feature space. This simplifies the projection of the original data $\mathbf{X}$ onto the principal components:

$$\mathbf{Y} = \mathbf{XV},$$

where $\mathbf{V}$ is the matrix of orthogonal eigenvectors. The orthogonality of $\mathbf{V}$ ensures that the transformation is numerically stable and efficient.

4. Interpretability

Orthogonal principal components make it easier to interpret the transformed data. Since each component is independent of the others, the contribution of each component to the overall variance can be analyzed separately. This is particularly useful for feature selection and data visualization.

5. Preservation of Distances

Orthogonal transformations preserve the Euclidean distances between data points. This means that the structure of the data is maintained in the lower-dimensional space, which is crucial for tasks like clustering and classification.

6. Computational Efficiency

Orthogonality reduces computational complexity. For example, inverting an orthogonal matrix $\mathbf{V}$ is straightforward since $\mathbf{V}^{-1} = \mathbf{V}^T$. This property is leveraged in algorithms like Singular Value Decomposition (SVD), which is often used to compute PCA.

Conclusion

The orthogonality of principal components is a fundamental property of PCA that ensures uncorrelated features, maximizes variance, simplifies computations, and preserves the structure of the data. These benefits make PCA a powerful and widely used tool for dimensionality reduction and data analysis.