Principal Component Analysis (PCA) Tutorial

Nayel Bettache

November 14, 2024

Contents

1 Introduction to PCA

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset while retaining as much information as possible. PCA finds new variables, called *principal components*, that are linear combinations of the original variables, capturing the directions of maximum variance in the data. This technique is widely used for data visualization, noise reduction, and as a preprocessing step for machine learning algorithms.

2 Mathematical Formulation of PCA

PCA aims to find the directions (principal components) that maximize the variance in the data. These components are the eigenvectors of the data's covariance matrix. The eigenvalues associated with these eigenvectors represent the amount of variance explained by each component.

2.1 Standardization of Data

PCA is sensitive to the scale of the variables, so we start by centering the data. Given a dataset $\mathbf{X}$ with n observations and p variables, we center the data by subtracting the mean of each variable.

Let $\mathbf{x}_i$ represent the i-th observation, and $\bar{\mathbf{x}}$ represent the mean vector:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

The centered data matrix $\mathbf{X}_{\text{centered}}$ is:

$$\mathbf{X}_{\text{centered}} = \mathbf{X} - \bar{\mathbf{x}}$$

2.2 Covariance Matrix

The covariance matrix, $\mathbf{\Sigma}$, is a $p \times p$ matrix that describes the variability of each variable and the relationships between pairs of variables. It is defined as:

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

2.3 Eigenvectors and Eigenvalues

PCA involves finding the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors represent the directions of maximum variance (principal components), and the eigenvalues indicate the amount of variance captured by each component.

Let $\mathbf{\Sigma}\mathbf{v} = \lambda\mathbf{v}$, where:

- $\mathbf{v}$ is an eigenvector of $\mathbf{\Sigma}$, representing a principal component.
- λ is the corresponding eigenvalue, indicating the variance along $\mathbf{v}$.

3 Geometric Interpretation of PCA

PCA projects the data onto a new coordinate system where the axes (principal components) represent the directions of maximum variance. The first principal component captures the most variance, the second component captures the next most, and so on. These components are orthogonal to each other, ensuring they capture unique patterns in the data.

3.1 Projection onto Principal Components

Let $\mathbf{V}$ be the matrix of eigenvectors of $\mathbf{\Sigma}$, ordered by their eigenvalues in decreasing order. We can project the data $\mathbf{X}_{\text{centered}}$ onto the principal components as follows:

$$\mathbf{Z} = \mathbf{X}_{\text{centered}}\mathbf{V}$$

where $\mathbf{Z}$ is the matrix of transformed data in the new coordinate system.

4 Worked Example

Consider the following dataset with two variables:

$$\mathbf{X} = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \\ 1.9 & 2.2 \\ 3.1 & 3.0 \\ 2.3 & 2.7 \\ 2.0 & 1.6 \\ 1.0 & 1.1 \\ 1.5 & 1.6 \\ 1.1 & 0.9 \end{bmatrix}$$

4.1 Step 1: Center the Data

Calculate the mean of each variable, subtract it from each observation, and obtain the centered data matrix.

4.2 Step 2: Covariance Matrix

Compute the covariance matrix $\mathbf{\Sigma}$ of the centered data.

4.3 Step 3: Eigenvalues and Eigenvectors

Find the eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}$. These represent the principal components.

4.4 Step 4: Project Data onto Principal Components

Transform the data by projecting it onto the principal components, reducing the dimensionality of the data.

5 Interpretation of PCA Results

The first principal component captures the largest possible variance. By examining the explained variance ratio, we can determine how much information each component retains. For example, if the first principal component explains 80% of the variance, and the second explains 15%, we could reduce dimensionality by projecting onto just the first component, retaining most of the data's information.

6 Practical Application: Using PCA for Visualization

PCA is especially useful for visualizing high-dimensional data in 2D or 3D. For example, in the Iris dataset, which has four features, PCA can reduce it to two principal components, allowing visualization in two dimensions.

7 PCA in Python

Below is sample Python code to perform PCA using `scikit-learn`:

```
from sklearn.decomposition import PCA
from sklearn.datasets import load_iris
import matplotlib.pyplot as plt

# Load the Iris dataset
data = load_iris()
X = data.data

# Apply PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Plot the transformed data
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=data.target, cmap=’viridis’)
plt.xlabel(’Principal Component 1’)
plt.ylabel(’Principal Component 2’)
plt.title(’PCA of Iris Dataset’)
plt.show()
```

8 Conclusion

PCA is a powerful tool for dimensionality reduction and visualization. By identifying directions of maximum variance, PCA helps capture the essence of the data in a smaller number of dimensions, often making it easier to analyze and visualize complex datasets.