Principal Component Analysis (PCA)

Short Report

1. Introduction

The data reduction technique Principal Component Analysis (PCA) stands as one of the most popular methods for dimensionality reduction. The main purpose of PCA is to transform complex datasets with numerous features into simpler forms which retain essential information. The algorithm identifies new axes called principle components which maximize data variation across different directions. The new axes maintain orthogonal relationships with each other while their ranking depends on the amount of data variance they explain.

The application of PCA precedes machine learning model implementation because it eliminates noise and redundant information while enabling better visualization of complex high-dimensional datasets.

2. Step-by-Step Process

Let's say we have a feature matrix $\mathbf{X}$ with n samples and d features:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Step 1: Center the Data

First, we subtract the mean of each feature to make sure the data is centered around zero:

$$\bar{\vec{x}} = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i, \quad \mathbf{X}_c = \mathbf{X} - \bar{\vec{x}}$$

This ensures that each column of $\mathbf{X}_c$ has a mean of zero.

Step 2: Compute the Covariance Matrix

Next, we calculate the covariance matrix, which captures how features vary with respect to each other:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

Each element C_{ij} represents the covariance between feature i and feature j.

Step 3: Eigen Decomposition

We then perform eigen decomposition of the covariance matrix:

$$\mathbf{C}\vec{w}_i = \lambda_i \vec{w}_i$$

Here, λ_i is the eigenvalue and $\vec{w}_i$ is the corresponding eigenvector. The eigenvectors represent the new directions (principal components), and the eigenvalues tell us how much variance is captured along each direction.

Step 4: Sort and Select Principal Components

We sort the eigenvalues in descending order:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$$

Then, we choose the top k eigenvectors that correspond to the largest eigenvalues. These k eigenvectors form the projection matrix $\mathbf{W}_k \in \mathbb{R}^{d\times k}$.

$$\mathbf{Z} = \mathbf{X}_c\mathbf{W}_k$$

The matrix $\mathbf{Z}$ represents the dataset transformed into the new k-dimensional subspace — this is our reduced representation.

Step 5: Choosing the Optimal k

To decide how many components to keep, we calculate the cumulative explained variance ratio (CEV):

$$\text{CEV}_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j}$$

We then plot CEV_k versus k and pick the value of k for which CEV_k is around 0.9 to 0.95. This means the selected components explain 90–95% of the total variance in the data.

3. Why PCA is Useful

PCA reduces redundant features and compresses data while still preserving the essential patterns. It's especially useful for:

- Visualizing high-dimensional data (e.g. projecting to 2D or 3D).
- Speeding up training of machine learning models.
- Removing noise and multicollinearity among features.

4. Advantages and Limitations

Advantages:

- Reduces complexity without much information loss.
- Makes visualization possible in fewer dimensions.
- Helps remove correlated or redundant features.

Limitations:

- PCA is a linear method — it may not work well if data is highly non-linear.
- The new principal components can be hard to interpret.
- It's sensitive to the scale of features, so data should be standardized first.

5. Conclusion

The analysis of complex high dimensional datasets becomes simpler through PCA which serves as a powerful tool for structural understanding. Dependable data analysis technique because it provides reliable results through its elegant mathematical framework.