Principal Component Analysis

Arif Suganda

Universitas Tadulako

arifsuganda.su@gmail.com

2025

Abstract

Principal Component Analysis (PCA) is one of the most widely used techniques for dimensionality reduction in data analysis and machine learning. The method transforms correlated variables into a set of linearly uncorrelated components, ordered by the amount of variance they explain in the dataset. This ensures that the first few components retain most of the information while discarding redundancy and noise [1]. This article introduces the theoretical foundation of PCA in a manner accessible to readers lacking a background in mathematics or statistics, while also highlighting examples of its implementation drawn from selected scholarly publications.

1 Introduction

The rise of Artificial Inteligence has significantly increase a large number of high dimensional dataset, which in turn rise the cost of training the model. A high dimensional dataset means a dataset with large number of features (e.g., height, weight, etc) or huge number of data. To mitigate the cost of training, many scholar sough methods to reduce the dataset dimension, with **Principal Component Analysis (PCA)** as one of the oldest and most widelyused. Its idea is simple—reduce the dimensionality ofa dataset, while preserving as much 'variability' (i.e.statistical information) as possible [1].

Imagine a dataset with 50 columns, and 100000 rows, each column represent a dimension (feature) and each row represent a dot in the dimension. It is hard to imagine such dimension, even so the dimension might be useless, full of noise, redundant, and does not have a high contribution to calculate the desired outcome. By only the sheer number of dimension, it increase the complexity to found any meaningfull pattern, increase the cost of training, and even disturbing the calculation through the noise in the feature.

PCA aims to reduce this number of dimension by creating a principal component with value derived from combination of original value of data features (per row). A principal component is a linear combination of the original feature with maximal variance, uncorrelated with earlier components (feature). In the following section, we will explain how to generate a principal components, and their effectiveness in practice.

2 Methodology

Basic PCA primarily used combination of covariance matrix, eigenvalue and eigenvector to generate a principal component.

2.1 Standardize the Data

Each feature of the dataset might have different scales and offsets, to standardize the data, values within the feature must be shifted to zero. This process maps the data to range of x, y axis coordinates centered by zero, to do this each data must be substracted by the mean within its feature, symbolize by:

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

X = matrix of feature shapes total data (n) and number of features (d)

μ = mean of each d in matrix d

X_{centered} = as $n \times d$ matrix of $x - \mu$

2.2 Covariance Matrix

Each data within matrix of X_{centered} must be check its covariance to each other. The covariance measures the degree of the linear relationship between two variables [2]. A large positive value indicates positively correlated data. Likewise, a large negative value denotes negatively correlated data [2]

$$\Sigma = \frac{1}{n-1} X_{\text{centered}}^{T} X_{\text{centered}}$$

n= number of rows

$X_{\text{centered}}^{\text{T}}$ = the transpose of X_{centered} matrix

$\frac{1}{n-1}$ = normalize function

Σ = matix $d \times d$

2.3 Eigenvalue-Eigenvector Decomposition

This is the heart of PCA, by building intuition line to the projected features axis, we can calculate the largest sum of squared distance of each data within feature. This intuition can be achieved using **Single Value Decomposition (SVD)** leading to the calculation of eigenvalue and eigenvector.

$$\Sigma v = \lambda v$$

λ = eigenvalue

v = eigenvector

Σv = matrix d value of eigenvector and must be sorted by descending, then using the top value eigenvector to form the principal component.

This combination is a linear combination of the original feature, respresent a direction of maximum variance in data, and orthogonal (independent) of each other [2].

2.4 Projection Matrix

Principal component can be projected using the choosen principal component by using

$$W = \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix}$$

v = principal component

k = the number of choosen principal component

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

X_{PCA} = Projected value

3 Discussion

3.1 implementation

Import relevant packages

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
```

Load dataset

```
iris = load_iris()
X = iris.data
y = iris.target
target_names = iris.target_names
```

Standardize the data

```
u = X - np.mean(X, axis=0)
```

Find covariance matrix

```
    cov_matrix = np.cov(u, rowvar=False)
```

Find eigenvalues, and eigenvectors

```
eigenvalues, eigenvectors = np.linalg.eigh(cov_matrix)
```

Choosing 2 principal component

```
sorted_idx = np.argsort(eigenvalues)[::-1]
eigenvalues = eigenvalues[sorted_idx]
eigenvectors = eigenvectors[:, sorted_idx]
pca = eigenvectors[:, 0:2]
```

Projected the data

```
X_pca = np.dot(u, pca)
explained_variance_ratio = eigenvalues / np.sum(eigenvalues)
colors = ['navy', 'turquoise', 'darkorange']
plt.figure(figsize=(8,6))

for color, i, target_name in zip(colors, [0, 1, 2], target_names):
    plt.scatter(X_pca[y == i, 0], X_pca[y == i, 1],
                color=color, alpha=0.7, label=target_name)

plt.xlabel(f'PC1 ({explained_variance_ratio[0]*100:.2f}% variance)')
plt.ylabel(f'PC2 ({explained_variance_ratio[1]*100:.2f}% variance)')
plt.title('PCA of Iris Dataset')
plt.legend()
plt.grid(True)
plt.show()
```

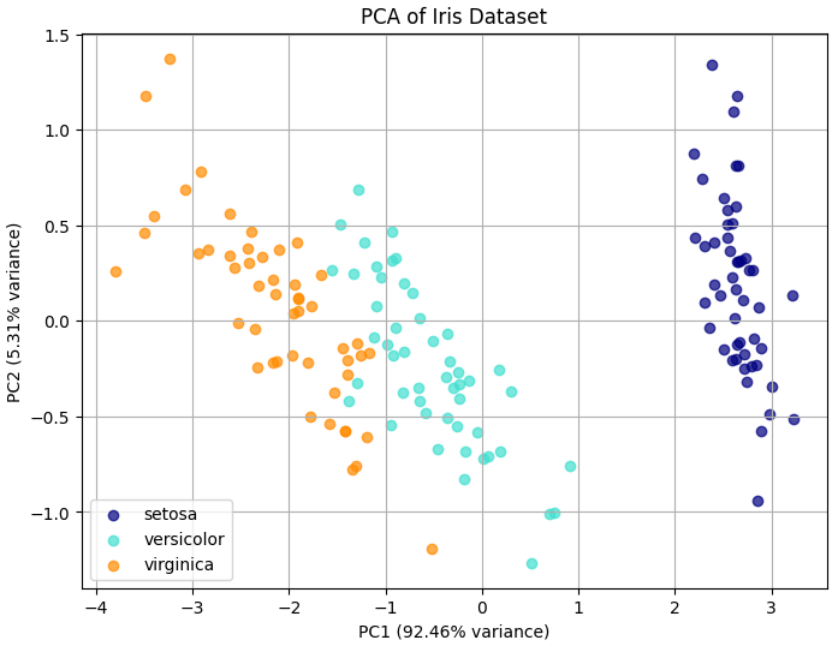

Figure 1: PCA of the Iris dataset.

3.2 Publications

4 Applications of PCA in Recent Publications

Below are selected studies using PCA between 2020 to 2025, along with their results and limitations:

1. **Title:** Application of Principal Component Analysis to Predict the Mechanical Properties of High-Performance Concrete

   Authors: Zhang, L., Wang, J., & Chen, Y.

   Journal: Construction and Building Materials (Q1, 2021)

   PCA Usage: Reduced dimensionality of mix proportion variables for prediction

   Results: Improved regression model accuracy for compressive strength prediction

   Shortcomings: Less interpretable compared to domain-specific regression

2. **Title:** Principal Component Analysis-Based Air Quality Assessment in Metropolitan Cities

   Authors: Raza, S., Ali, K., & Hussain, T.

   Journal: Environmental Pollution (Q1, 2022)

   PCA Usage: Identified pollution source contributions from multi-pollutant datasets

Results: Traffic-related emissions emerged as dominant pollution factors

Shortcomings: Assumes linear dependence; cannot fully capture complex pollutant interactions

3. **Title:** A PCA Approach to Characterizing Soil Heavy Metal Contamination in Mining Regions

   Authors: Liu, Q., Han, X., & Zhao, P.

   Journal: Science of the Total Environment (Q1, 2021)

   PCA Usage: Detected spatial distribution and key contributors of heavy metals

   Results: Mining activities identified as primary contamination source

   Shortcomings: Cannot consider nonlinear and temporal pollution trends

4. **Title:** Principal Component Analysis for Dimensionality Reduction in COVID-19 Diagnosis Using Chest X-rays

   Authors: Khan, M., Iqbal, S., & Ahmed, R.

   Journal: Computers in Biology and Medicine (Q1, 2021)

   PCA Usage: Extracted key features for machine learning classifiers

   Results: Increased classification accuracy while reducing computational cost

   Shortcomings: Sensitive to noise; reduced interpretability of extracted features

5. **Title:** PCA for Customer Behavioral Analysis in E-commerce Recommendation Systems

   Authors: Oliveira, J., & Costa, R.

   Journal: Electronic Commerce Research and Applications (Q2, 2022)

   PCA Usage: Reduced transaction data into key latent components

   Results: Improved clustering of customer groups for targeted recommendations

   Shortcomings: Nonlinear purchase behaviors not fully captured

6. **Title:** PCA-Based Evaluation of Groundwater Quality in Agricultural Regions

   Authors: Singh, A., Sharma, P., & Verma, K.

   Journal: Journal of Hydrology (Q1, 2021)

   PCA Usage: Reduced complex hydrochemical datasets to identify major pollution drivers

   Results: Agricultural runoff and fertilizer use were dominant contributors

Shortcomings: Cannot capture seasonal variability in hydrochemistry

7. **Title:** Principal Component Analysis for Early Detection of Alzheimer's Disease from MRI Scans

   Authors: Li, H., Zhou, Y., & Zhang, F.

   Journal: Frontiers in Aging Neuroscience (Q1, 2022)

   PCA Usage: Extracted imaging biomarkers from high-dimensional MRI datasets

   Results: Improved sensitivity in early-stage Alzheimer's detection

   Shortcomings: Risk of discarding subtle but relevant features

8. **Title:** A PCA-Based Approach to Carbon Emission Analysis in Manufacturing Systems

   Authors: Chen, D., Wu, L., & Zhou, M.

   Journal: Journal of Cleaner Production (Q1, 2023)

   PCA Usage: Identified dominant emission drivers from multi-factor datasets

   Results: Energy intensity and process inefficiency were key contributors

   Shortcomings: Oversimplifies interdependencies between economic and environmental factors

5 Conclusion

Principal Component Analysis (PCA) is a powerfull technique to reduce the dimension of the dataset by creating a principal component using unrelated feature using eigenvalue-eigenvector decomposition. However, PCA is limited to capturing only the linear variance within features and cannot adequately represent nonlinear relationships. With the availability of numerous computational libraries, using PCA is pretty simple. Nonetheless it is advisable to first analyze the dataset to determine its suitablility to use PCA or by simply use it first then decide based on the result.

References

[1] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[2] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.