Chapter 3: Principal Component Analysis (PCA)

Latreche Sara

Contents

Introduction to PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique. Its goal is to reduce the number of features in a dataset while preserving as much variability as possible. PCA achieves this by identifying directions—called **principal components**—along which the data varies the most.

The technique is particularly useful when dealing with high-dimensional data, where visualization and modeling can become difficult. By projecting data onto a lower-dimensional space, PCA helps uncover patterns, compress data, and denoise input features.

Key Use Case

PCA is often used as a preprocessing step before applying machine learning algorithms, especially when features are highly correlated.

1 Motivation for Dimensionality Reduction

Question: Why might we want to reduce the number of features?

Some possible reasons include:

- Reducing noise and irrelevant information
- Lowering computational cost
- Avoiding the curse of dimensionality
- Enhancing interpretability

Study Question: What happens when you project high-dimensional data onto a lower-dimensional subspace?

2 PCA Intuition

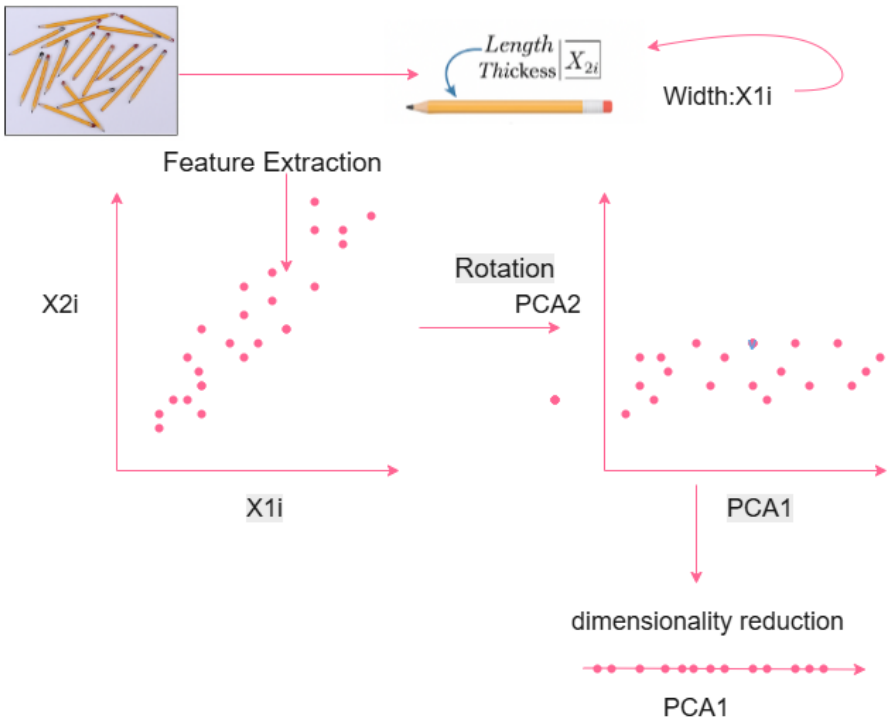

Figure 1: Illustration of PCA using pencils: Each pencil has two features (length and width). PCA finds the main direction of variance and reduces dimensions.

To intuitively understand PCA, imagine a collection of pencils scattered across a table. Each pencil can be described using two features: its **length** and its **thickness**. We represent each pencil as a point in a 2D feature space.

Step 1: Feature Extraction
We measure the length and thickness of each pencil and plot them. Often, longer pencils may also be thicker, creating a diagonal distribution of points — indicating correlation between features.

Step 2: Rotation (PCA)
PCA identifies the directions in which the data varies the most. These direc-

tions are called *principal components*. PCA1 (the first principal component) is the direction with the greatest variance. PCA2 is orthogonal to PCA1 and captures any remaining variation.

Step 3: Dimensionality Reduction
Since PCA1 already captures most of the important variation, we can project our data onto it and discard PCA2. Now, instead of describing each pencil with two numbers (length and thickness), we use just one value — the position along PCA1. This results in simpler data with minimal loss of information.

PCA reduces dimensionality while preserving the essential structure of the data. It removes redundancy, simplifies patterns, and is particularly useful before visualization or further machine learning tasks.

PCA Toolbox (Essence)

PCA rotates the data to find new axes (principal components) that capture the most variance. Keep the top ones, drop the rest — fewer dimensions, same essence.

3 Correlation, Covariance, and Standardization

Consider a dataset with Q objects and N features, arranged as in Table 1.

	Object 1	Object 2	$\cdots$ Object Q
Feature 1	5.4	2.4	12.3
Feature 2	7.5	3.5	10.3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Feature N	8.3	1.4	14.2

Table 1: Typical data layout for PCA analysis.

Each feature corresponds to a random variable, so each object is a point in an N-dimensional feature space. Since N is usually large, visualization is chal-

lenging. Dimensionality reduction techniques like PCA help by projecting the data onto lower dimensions.

Before analysis, features are often *standardized* by subtracting the mean μ_X and dividing by the standard deviation σ_X:

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}.$$

This transformation ensures each feature has zero mean and unit variance, making them comparable regardless of original units.

To quantify relationships between two random variables X_1 and X_2, we use three key measures:

- **Correlation:** $R_{12} = E[X_1 X_2]$ is the expected product of the two variables. It gives a raw measure of how the variables vary together but does not account for their means or scales.
- **Covariance:**

  $$\mathrm{Cov}(X_1, X_2) = E\big[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})\big].$$

  This measures how two variables jointly vary after centering (removing their means). Positive covariance indicates they tend to increase or decrease together, while negative means they vary in opposite directions. Its magnitude depends on the units of the variables.
- **Pearson correlation coefficient:**

  $$C_{12} = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = E[\tilde{X}_1 \tilde{X}_2],$$

  which normalizes covariance by the standard deviations, resulting in a dimensionless measure bounded between -1 and 1. It captures the strength and direction of the linear relationship independent of scale. Values close to ± 1 imply a strong linear relationship; values near zero imply weak or no linear association.

These concepts generalize to covariance and correlation matrices for all N features, forming the mathematical foundation for many multivariate analysis techniques, including PCA.

Numerical Example: Correlation, Covariance, and Standardization

Consider two features X_1 and X_2 measured over three objects, as shown below:

	Object 1	Object 2	Object 3
X_1	2	4	6
X_2	1	3	5

- **Compute the means:**

$$\mu_{X_1} = \frac{2+4+6}{3} = 4, \quad \mu_{X_2} = \frac{1+3+5}{3} = 3.$$

- **Center the data by subtracting the mean:**

$$X_1 - \mu_{X_1} = [-2, 0, 2], \quad X_2 - \mu_{X_2} = [-2, 0, 2].$$

- **Calculate covariance:**

$$\text{Cov}(X_1, X_2) = \frac{1}{3}\sum_{i=1}^{3}(X_{1i} - \mu_{X_1})(X_{2i} - \mu_{X_2}) = \frac{8}{3} \approx 2.67.$$

- **Calculate standard deviations:**

$$\sigma_{X_1} = \sqrt{\frac{1}{3}(4+0+4)} = \sqrt{\frac{8}{3}} \approx 1.63,$$

$$\sigma_{X_2} = \sqrt{\frac{1}{3}(4+0+4)} = \sqrt{\frac{8}{3}} \approx 1.63.$$

- **Calculate Pearson correlation coefficient:**

$$C_{12} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{2.67}{1.63 \times 1.63} = 1.$$

This means X_1 and X_2 have a perfect positive linear relationship.

Standardized Data:

$$\tilde{X_1} = \frac{X_1 - \mu_{X_1}}{\sigma_{X_1}} = \left[-\frac{2}{1.63}, 0, \frac{2}{1.63}\right] \approx [-1.22, 0, 1.22],$$
$$\tilde{X_2} = \frac{X_2 - \mu_{X_2}}{\sigma_{X_2}} = \left[-\frac{2}{1.63}, 0, \frac{2}{1.63}\right] \approx [-1.22, 0, 1.22].$$

	Object 1	Object 2	Object 3
$\tilde{X_1}$	-1.22	0	1.22
$\tilde{X_2}$	-1.22	0	1.22

Notice how both features are now centered at zero mean and scaled to unit variance, allowing meaningful comparison.

4 Principal Component Analysis

In this section, we present the mathematical formulation of Principal Component Analysis (PCA). To simplify the discussion, we integrate the conceptual framework outlined in the introduction with the statistical concepts introduced previously.

Let the dataset be represented by a matrix $\mathbf{X} \in \mathbb{R}^{N \times Q}$, where each column corresponds to an object (or individual), and each row i ($1 \leq i \leq N$) corresponds to a particular measurement or feature vector $\mathbf{X}_i$. The values measured for the j-th object are represented as $\mathbf{X}_j$.

PCA is a linear transformation and can be expressed in matrix form as:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}, \tag{1}$$

where $\mathbf{W}$ is the transformation matrix, and $\mathbf{Y}$ is the transformed data matrix. Our goal is to determine the matrix $\mathbf{W}$ that maps the original features into a new coordinate system.

4.1 Centering the Data

To begin, we compute the empirical mean vector $\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2}, \ldots, \mu_{X_N})^T$ of dimension $N \times 1$, where

$$\mu_{X_i} = \frac{1}{Q} \sum_{j=1}^{Q} X_{ij}. \tag{2}$$

We center the data by subtracting the mean from each measurement:

$$\hat{X}_{ij} = X_{ij} - \mu_{X_i}. \tag{3}$$

Let $\hat{\mathbf{X}}$ denote the mean-centered data matrix.

4.2 Covariance Matrix and Eigen Decomposition

The covariance matrix of the centered data is given by:

$$\mathbf{K} = \mathrm{Cov}(\hat{\mathbf{X}}) = \frac{1}{Q-1} \hat{\mathbf{X}} \hat{\mathbf{X}}^T. \tag{4}$$

We then perform eigen decomposition of the covariance matrix $\mathbf{K}$ to obtain eigenvalues $\lambda_i \geq 0$ and corresponding eigenvectors $\mathbf{v}_i$ for $1 \leq i \leq N$. These eigenvectors form the rows of the transformation matrix:

$$\mathbf{W} = \begin{bmatrix} \leftarrow \mathbf{v}_1^T \rightarrow \\ \vdots \\ \leftarrow \mathbf{v}_N^T \rightarrow \end{bmatrix}. \tag{5}$$

The PCA projection of object j is then given by:

$$\mathbf{Y}_j = \mathbf{W} \mathbf{X}_j. \tag{6}$$

4.3 Variance Preservation and Dimensionality Reduction

The total variance of the transformed data is:

$$S = \sum_{i=1}^{N} \sigma_{Y_i}^2 = \sum_{i=1}^{N} \lambda_i. \tag{7}$$

Since PCA is a rotation of the coordinate system, the total variance is preserved:

$$\text{Var}(\mathbf{X}) = \text{Var}(\mathbf{Y}). \tag{8}$$

To reduce the dimensionality from N to $M < N$, we retain the first M eigenvectors corresponding to the largest eigenvalues. The variance preserved by this projection is:

$$S_c = \sum_{i=1}^{M} \lambda_i, \tag{9}$$

and the ratio of preserved variance is:

$$G = 100 \cdot \frac{S_c}{S}. \tag{10}$$

This criterion allows us to choose M such that a desired percentage of the total variance (e.g., $G \approx 70\%$) is retained.

Several methods exist to determine the optimal value of M, including probabilistic PCA [**?**], Bayesian treatments [**?**], and generalization error minimization [**?**].

It is important to note that not all datasets guarantee a low-dimensional representation with high variance preservation. The effectiveness of PCA in reducing dimensionality while preserving structure depends on the correlation between features in the dataset.

Numerical Example: PCA Step-by-Step

Dataset: Consider the following matrix with $N = 2$ features and $Q = 4$ individuals:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & 6 & 8 \\ 1 & 3 & 5 & 7 \end{bmatrix}$$

1. Center the Data

$$\mu_{X_1} = 5, \quad \mu_{X_2} = 4 \Rightarrow \hat{\mathbf{X}} = \begin{bmatrix} -3 & -1 & 1 & 3 \\ -3 & -1 & 1 & 3 \end{bmatrix}$$

2. Covariance Matrix

$$\mathbf{K} = \frac{1}{3}\hat{\mathbf{X}}\hat{\mathbf{X}}^T = \frac{1}{3}\begin{bmatrix} 20 & 20 \\ 20 & 20 \end{bmatrix} = \begin{bmatrix} 6.67 & 6.67 \\ 6.67 & 6.67 \end{bmatrix}$$

3. Eigenvalues and Eigenvectors

$$\lambda_1 = 13.34, \quad \lambda_2 = 0 \mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

4. Projected Data (PCA Transformation)

$$\mathbf{W} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{Y} = \mathbf{W}\hat{\mathbf{X}} = \begin{bmatrix} -4.24 & -1.41 & 1.41 & 4.24 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Conclusion: The second principal component has zero variance. All data points lie on a line in 2D space, so PCA successfully reduces the dimension to 1D.

Conceptual Questions for Deeper Understanding

1. Why is the second eigenvalue zero in this example?
2. How is the first principal component related to the original axes?
3. What would the PCA results look like if the data were uncorrelated?
4. Can PCA always reduce the dimension to 1D? Under what conditions?
5. Try plotting the original points and the first principal component—what do you observe?

Exercise: Choosing the Number of Principal Components

Context: Suppose PCA was performed on a dataset with $N = 5$ features. After computing the covariance matrix and performing eigen decomposition, we obtain the following sorted eigenvalues:

$$\lambda_1 = 5.2, \quad \lambda_2 = 3.1, \quad \lambda_3 = 1.1, \quad \lambda_4 = 0.4, \quad \lambda_5 = 0.2$$

Total Variance: $S = \sum_{i=1}^{5} \lambda_i = 10.0$

Q1. Retained Variance for $M = 1$:

$$S_c = \lambda_1 = 5.2, \quad G = \frac{5.2}{10.0} \cdot 100 = 52\%$$

Q2. Retained Variance for $M = 2$:

$$S_c = \lambda_1 + \lambda_2 = 8.3, \quad G = 83\%$$

Q3. Retained Variance for $M = 3$:

$$S_c = \lambda_1 + \lambda_2 + \lambda_3 = 9.4, \quad G = 94\%$$

Conclusion: If the goal is to preserve at least 90% of the variance, then $M = 3$ is sufficient.

Questions to Reflect On

1. What are the advantages of using fewer components when G is still high?
2. Why might we choose $M = 2$ even though $M = 3$ gives higher G?
3. What could be the risks of choosing a very small M?
4. How does this method relate to the concept of a scree plot?
5. Can we always trust variance alone to select the best M?

5 Other Aspects of PCA

Principal Component Analysis (PCA) is not only a dimensionality reduction technique but also offers rich geometric and optimization interpretations. In this section, we elaborate on two important aspects: the geometric interpretation of PCA and its formulation as a variance maximization problem using the Lagrangian method.

5.1 Geometric Interpretation of PCA

PCA can be viewed as a geometric transformation of the data. It rotates the original coordinate system so that the new axes (principal components) align with the directions of maximum variance in the data. These axes are orthogonal (i.e., perpendicular), and each subsequent axis accounts for the maximum remaining variance under the orthogonality constraint.

Given a dataset represented by a matrix $X \in \mathbb{R}^{N \times Q}$, where each column is a data sample, PCA finds a set of orthonormal vectors (the principal components) $\{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_N\}$ such that:

- $\vec{w}_1$ points in the direction of maximum variance,
- $\vec{w}_2$ points in the direction of maximum remaining variance orthogonal to $\vec{w}_1$,
- and so on.

PCA effectively projects the data onto these new axes. Geometrically, this projection can be interpreted as fitting an ellipsoid to the data cloud where the principal axes of the ellipsoid correspond to the eigenvectors of the covariance matrix. The lengths of these axes correspond to the square roots of the respective eigenvalues, indicating the variance along each principal component.

This perspective makes PCA a powerful tool for understanding intrinsic data structure, especially in high dimensions, as the principal directions capture the most significant modes of variation in the data.

5.2 PCA as a Rotation

A key property of PCA is that it can be seen as a rotation of the data matrix. Suppose X is centered (i.e., zero mean), and let $W \in \mathbb{R}^{N\times N}$ be an orthogonal matrix such that $W^T = W^{-1}$. The rotated data matrix is then:

$$Y = WX. \tag{11}$$

The covariance matrix of Y is given by:

$$\mathrm{Cov}(Y) = \frac{1}{Q-1}(Y - \boldsymbol{\mu}_Y \mathbf{h}^T)(Y - \boldsymbol{\mu}_Y \mathbf{h}^T)^T, \tag{12}$$

where $\boldsymbol{\mu}_Y = W\boldsymbol{\mu}_X$ is the mean of the rotated data and $\mathbf{h} \in \mathbb{R}^{1\times Q}$ is a vector of ones.

Substituting the expression for Y and simplifying, we get:

$$\mathrm{Cov}(Y) = \frac{1}{Q-1}(WX - W\boldsymbol{\mu}_X \mathbf{h}^T)(WX - W\boldsymbol{\mu}_X \mathbf{h}^T)^T \tag{13}$$

$$= \frac{1}{Q-1}W(X - \boldsymbol{\mu}_X \mathbf{h}^T)(X - \boldsymbol{\mu}_X \mathbf{h}^T)^T W^T \tag{14}$$

$$= W\,\mathrm{Cov}(X)\,W^T. \tag{15}$$

Since W is orthogonal, we can also write:

$$\mathrm{Cov}(Y) = W\,\mathrm{Cov}(X)\,W^{-1}. \tag{16}$$

This equation shows that rotation by an orthogonal matrix preserves the eigenvalues of the covariance matrix. If we choose W to be the matrix of eigenvectors of $\mathrm{Cov}(X)$, then $\mathrm{Cov}(Y)$ is diagonal — each diagonal element corresponding to the variance along a principal component.

5.3 Variance Maximization via Lagrangian Derivation

PCA can also be derived as an optimization problem where we seek directions that maximize the variance of projected data. Let us define the transformed data as:

$$Y = WX, \tag{17}$$

where each row of W corresponds to a new basis vector (principal component), and we want each projected component $Y_i = \vec{w}_i X$ to have maximal variance.

The variance of Y_i is:

$$\mathrm{Var}(Y_i) = \vec{w}_i \,\mathrm{Cov}(X)\, \vec{w}_i^T. \tag{18}$$

To avoid the trivial solution $\vec{w}_i \to \infty$, we impose the constraint $\|\vec{w}_i\| = 1$. This leads us to a constrained optimization problem using a Lagrangian:

$$\mathcal{L}(\vec{w}_i, \lambda_i) = \vec{w}_i \,\mathrm{Cov}(X)\, \vec{w}_i^T - \lambda_i(\vec{w}_i \vec{w}_i^T - 1). \tag{19}$$

Taking the derivative with respect to $\vec{w}_i$ and setting it to zero:

$$\frac{d\mathcal{L}}{d\vec{w}_i} = 2\,\mathrm{Cov}(X)\, \vec{w}_i^T - 2\lambda_i \vec{w}_i^T = 0 \tag{20}$$

$$\Rightarrow \mathrm{Cov}(X)\, \vec{w}_i^T = \lambda_i \vec{w}_i^T. \tag{21}$$

Hence, $\vec{w}_i^T$ is an eigenvector of the covariance matrix, and λ_i is the corresponding eigenvalue. Since:

$$\mathrm{Var}(Y_i) = \vec{w}_i \,\mathrm{Cov}(X)\, \vec{w}_i^T = \lambda_i, \tag{22}$$

we conclude that the principal component corresponding to the largest variance is given by the eigenvector associated with the largest eigenvalue of $\mathrm{Cov}(X)$.

Stacking the top M eigenvectors as rows of $W \in \mathbb{R}^{M \times N}$, we obtain the PCA transformation matrix:

$$Y = WX, \tag{23}$$

where the projection $Y \in \mathbb{R}^{M \times Q}$ maximally preserves variance in M dimensions.

This derivation highlights how PCA aligns the new basis vectors with the directions of greatest variance in the data while maintaining orthogonality between them.

6 Comprehension Questions

Comprehension Questions

- What property must the rotation matrix W satisfy and why?
- Explain the effect of rotation on the covariance matrix.
- Why do the eigenvectors of the covariance matrix correspond to directions of maximal variance?
- How can eigenvalues help in deciding the number of principal components to retain?

7 PCA Summary Algorithm

PCA Algorithm Summary and Pseudocode

Summary:

1. Center the data by subtracting the mean of each feature.
2. Calculate the covariance matrix Σ.
3. Compute eigenvalues and eigenvectors of Σ.
4. Sort eigenvectors by decreasing eigenvalues.
5. Select the top k eigenvectors to form the projection matrix U_{reduce}.
6. Project the data onto the reduced space: $Z = U_{\text{reduce}}^T X$.

Pseudocode:

procedure PCA(X, k)
    Input: Data matrix $X \in \mathbb{R}^{N \times Q}$, number of components k
    Output: Reduced data $Z \in \mathbb{R}^{k \times Q}$, projection matrix $U_{\text{reduce}} \in \mathbb{R}^{N \times k}$, eigenvalues λ

Compute the mean vector: $\mu = \frac{1}{Q}\sum_{i=1}^{Q} X_i$
Center the data: $X \leftarrow X - \mu \mathbf{1}^T$
Compute covariance matrix: $\Sigma = \frac{1}{Q-1} X X^T$
Compute eigenvalues $\{\lambda_i\}$ and eigenvectors $\{u_i\}$ of Σ
Sort eigenvectors u_i by decreasing eigenvalues λ_i
Form projection matrix with top k eigenvectors: $U_{\text{reduce}} = [u_1, u_2, \ldots, u_k]$
Project the data: $Z = U_{\text{reduce}}^T X$
return $Z, U_{\text{reduce}}, \{\lambda_i\}$
end procedure

8 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised projection method used for categorized data, where each sample belongs to a specific class. Unlike PCA, which is unsupervised and focuses on maximizing variance, LDA aims to maximize the separation between different classes by considering their scatter distances.

Suppose the dataset contains C distinct classes. For the j-th class C_j, the scatter matrix is defined as:

$$S_j = \sum_{i \in C_j} (\mathbf{X}_i - \boldsymbol{\mu}_{C_j})(\mathbf{X}_i - \boldsymbol{\mu}_{C_j})^T,$$

where $\mathbf{X}_i$ is the feature vector for sample i, and $\boldsymbol{\mu}_{C_j}$ is the mean vector of class C_j.

The *within-class scatter matrix* (intra-group scatter) sums the scatter matrices of all classes:

$$S_{\text{intra}} = \sum_{j=1}^{C} S_j.$$

The *between-class scatter matrix* measures the dispersion of the class means

relative to the overall mean $\boldsymbol{\mu}_X$:

$$S_{\text{inter}} = \sum_{j=1}^{C} Q_j (\boldsymbol{\mu}_{C_j} - \boldsymbol{\mu}_X)(\boldsymbol{\mu}_{C_j} - \boldsymbol{\mu}_X)^T,$$

where Q_j is the number of samples in class C_j.

LDA finds the projection matrix by maximizing the following criterion:

$$S = S_{\text{intra}}^{-1} S_{\text{inter}}.$$

The eigenvectors of matrix S correspond to directions that maximize class separability. The LDA algorithm follows the same steps as PCA, but uses matrix S in place of the covariance matrix.

—

LDA is widely used for dimensionality reduction when class labels are available, enhancing classification performance by emphasizing class separability.

9 Limitations and Considerations

- PCA assumes linear relationships
- PCA is sensitive to scaling
- The principal components are linear combinations—not necessarily meaningful features
- PCA may discard small-variance features that are actually important

10 Applications of PCA Across Domains

Principal Component Analysis (PCA) is widely regarded as a powerful tool for reducing dimensionality, extracting structure, and revealing hidden patterns in data. Its mathematical simplicity and interpretability make it applicable across a broad range of disciplines. Below we outline a selection of domains where PCA has had substantial impact.

10.1 Economics and Development Studies

In economic analysis, PCA has been employed to synthesize multiple macroeconomic indicators into unified indices for evaluating national and regional development. For example, PCA has been used to construct a Sustainable Development Index for India, aggregating variables such as income, education, and health outcomes. Similarly, it has been applied in China to quantify the coordination of regional development, and in Ghana to examine differences in living conditions between urban and rural areas. These applications demonstrate PCA's utility in reducing complex social and economic data into principal components that represent underlying developmental gradients.

10.2 Scientometrics and Research Evaluation

In the study of scientific output and impact, PCA has been instrumental in disentangling the components of widely-used university rankings such as the Academic Ranking of World Universities (ARWU). It was shown that over 80% of the variance in ARWU's six metrics could be explained by just two principal components, highlighting how a small set of latent dimensions—such as institutional size and research excellence—govern ranking behavior. PCA has also been used to classify dozens of research impact metrics into citation-based, usage-based, and social-media-based dimensions, allowing for a more structured understanding of scientific influence. More recently, functional PCA has been used to identify "evergreen" papers, which continue to receive citations over long periods without decay, thereby revealing novel citation patterns.

10.3 Physics and Physical Chemistry

In physics, PCA has emerged as a data-driven technique to extract physical features from high-dimensional systems. It has been applied in condensed matter physics for phase recognition, including identifying topological and quantum phases without prior knowledge of order parameters. In nuclear physics, PCA has helped in the classification of nuclear structures by reducing hundreds of correlated variables—such as energy levels and transition

strengths—into a few dominant patterns. PCA has also been applied in physical chemistry and fluid dynamics to study particle trajectories in Brownian motion. Furthermore, the quantum analog of PCA (qPCA) has been proposed to estimate the eigenvalues and eigenvectors of quantum states, opening the door for its application in quantum state tomography and quantum machine learning.

10.4 Astronomy and Astrophysics

Astronomical datasets often contain vast amounts of spectral and imaging data. PCA has proven especially valuable in preprocessing tasks like spectral compression. For example, in stellar classification tasks, PCA has been used to reduce the dimensionality of input spectra prior to feeding them into artificial neural networks, improving both convergence and performance. In extragalactic astronomy, PCA has facilitated the classification of Seyfert galaxies, where spectra of type 1 objects are often well-represented by a single principal component, while type 2 objects require multiple components. Additionally, PCA has been used to enhance pulsar signal detection from waterfall diagrams and to identify underlying trends in large-scale photometric surveys.

10.5 Geography, Spatial Analysis, and Genetics

PCA has also found applications in geographical and spatial analysis. In innovation studies, PCA has been applied to patent data and company locations across Europe to detect cross-country patterns of technological development. These principal components often reflect not only national differences but also spatial regularities that transcend country borders. A particularly striking example of PCA's explanatory power is found in population genetics: principal components derived from genome-wide SNP data align closely with geographic coordinates, demonstrating how human genetic variation reflects migration and historical population structure.

10.6 Summary

Across all these domains—ranging from economics and scientometrics to quantum physics and astrophysics—PCA serves as a fundamental tool for uncovering structure in high-dimensional data. Its ability to distill complexity into interpretable dimensions makes it a cornerstone technique in both exploratory analysis and predictive modeling.

Conclusion

PCA is a powerful unsupervised learning technique used to reduce dimensionality while preserving the most important structure in the data. By identifying principal components, it enables more efficient storage, visualization, and modeling of high-dimensional datasets.

Understanding the mathematics behind PCA not only strengthens your grasp of linear algebra and optimization, but also provides a foundation for more advanced techniques like kernel PCA, t-SNE, and autoencoders.

References

[1] Karl Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.

[2] I.T. Jolliffe. *Principal Component Analysis: A Natural Approach to Data Exploration*. Springer Series in Statistics, 2nd edition, 2002.

[3] Andrew Ng. CS229 Lecture Notes: Principal Component Analysis. Stanford University. Available at: `https://cs229.stanford.edu/notes2022fall/cs229-notes10.pdf`

[4] Jonathon Shlens. A Tutorial on Principal Component Analysis. arXiv:1404.1100 [cs.LG], 2014. `https://arxiv.org/abs/1404.1100`