Principal Components Analysis

Qianning Zhang, Zhuoyang Lyu, Yinan Wu

What is Principal Component Analysis?

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of large datasets while preserving as much variability as possible.

It works by:

Finding new variables called principal components, which are linear combinations of the original variables.

These components are uncorrelated and ordered so that the first few capture most of the variation in the data.

What Problems Can PCA Solve in the Real World?

1. Too Many Variables (High Dimensionality)

The problem with having high dimensionality is the Complex datasets with hundreds of features are hard to interpret or visualize.

What PCA do is it Reduce to 2 or 3 principal components for visualization or analysis without much information loss.

2. Correlated Features

The problem is that Many variables are highly correlated (multicollinearity), which can confuse models.

PCA Solution: Produces uncorrelated components, improving model performance.

3. Noisy or Redundant Data

In this case, Noise or irrelevant features dilute meaningful patterns.

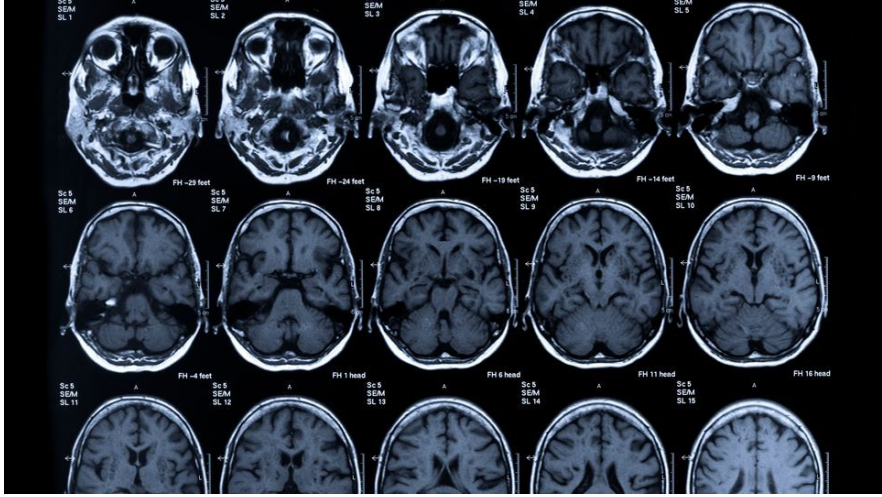
PCA Solution: Keeps the most informative features, filtering out noise.

4. Slow Computation in Machine Learning

Which happen when Algorithms are slow or overfit due to too many features.

PCA Solution: Speeds up training and reduces overfitting by trimming feature space.

Real-Life Applications of PCA



1. Genetics & Genomics

PCA is used to analyze gene expression data and identify population structures in genome-wide studies.

Helps reduce noise in datasets with thousands of genes per sample.

2. Neuroscience

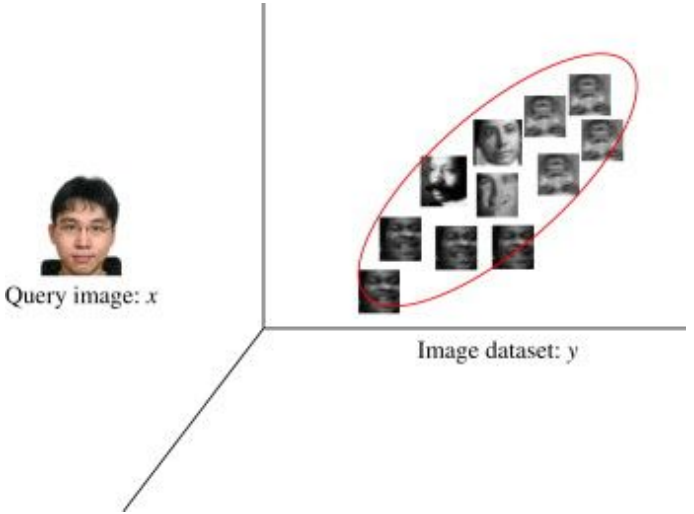
In brain imaging, PCA helps reduce dimensionality and identify patterns in brain activity signals.

Assists in identifying regions of interest or reducing artifacts.

3. Computer Vision & Image Compression

PCA compresses high-resolution images by keeping components that retain most visual information.

Used in facial recognition systems (e.g., Eigenfaces).



4. Finance

Analysts use PCA to identify underlying factors that drive asset returns or market movements.

Helps reduce hundreds of financial indicators into a few interpretable components.

5. Marketing & Customer Segmentation

PCA is applied to customer behavior data (purchases, clicks) to group similar customers for targeted marketing.

Mathematical Explanation of PCA

The objective is to find directions (vectors) that maximize variance. These directions are orthogonal and forms a low-dimensional representation with minimal information loss.

Initially we have a matrix with $X \in R^{(n \times d)}$, n is number of data points d is the number of features. And each row of X is a data point in R^d.

1

Compute the mean of each column $z = \frac{value - mean}{standard\ deviation}$

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^d$ (the mean vector).

subtract the mean for each member in their particular column.

$$\tilde{X} = X - \mathbf{1}\mu^T$$

This step is called mean-centering, making the mean for every column 0, which means the data is now centered around the origin in multi-dimensional space.

2

Compute how features vary with respect to each other. (covariance matrix)

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

The covariance matrix tells you how variables in a dataset vary together.
It captures both the variance of individual variables and the relationships (covariances) between different variables.

3

Decompose the covariance matrix into its eigenvalues and eigenvectors.

$$C = Q\Lambda Q^T$$

Q is a matrix of orthonormal eigenvectors of C.

Each column qi is a principal component direction which is a direction in your dataset's space along which the data varies the most.

Λ is a diagonal matrix of eigenvalues λ1,λ2,...,λd.

Each λi≥0 represents how much variance is explained along the direction qi

Eigenvectors give the directions of principal components.

Eigenvalues tell you how much variance is captured along each direction.

4

Choose the top k eigenvectors corresponding to the largest eigenvalues.

$$Q_k = [q_1\ q_2\ \dots\ q_k] \in \mathbb{R}^{d \times k}$$

This is because we want the subspace that captures the most variance in the data. Only this way we could reduce the information loss.

5

Transform the original data into the new k-dimensional basis.

$$Z = \tilde{X} Q_k \in \mathbb{R}^{n \times k}$$

Rows of Z are the low-dimensional representations of the original data.

This projection minimizes reconstruction error in the least squares sense.

6 (optional)

Map reduced data back to original space to see approximation.

$$\hat{X} = ZQ_k^T + \mu^T$$

This is taking the reduced data and projecting it back into the original d-dimensional space to reconstruct an approximation of the original data. This is helpful when you want to see how much information was lost by reducing the dimensionality.

Visualization of PCA

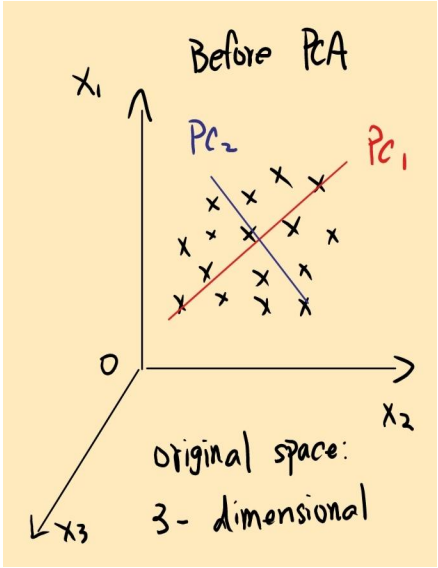
What is PCA visualization

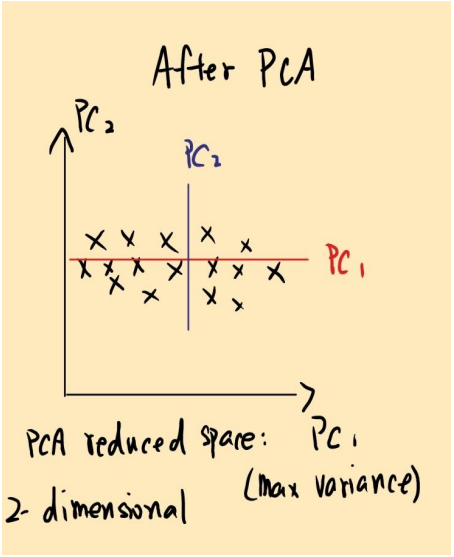
The process of reducing high-dimensional data into 2D or 3D using principal components, so we can see the structure more clearly. By capturing the most variances and the second most variances, PC1 and PC2, and making them be x axis and y axis, we can project high-dimensional data onto new axes and then create a 2D scatterplot. This helps reveal pattern like clusters or separations

PC1: The direction in the data space where the data varies the most (most variance)

PC2: The second most various direction, orthogonal to PC1

Left graph: before PCA
-It has raw data raw data in three dimensions, x1, x2 and x3.
-The data is scattered across all three axes (even though we only see a 2D projection), making it harder to interpret or visualize
-PC_1 (red line): the direction with the maximum variance
-PC_2 (blue line)**:** the second principal component, orthogonal to PC_1.





Right graph: after PCA
-The data has been transformed into a 2D space instead of 3D defined by PC_1 and PC_2
-PC1 is new x axis, PC2 is new y axis
-The spread of the data is greatest along PC_1 (the direction with the most variance), which is why PCA retains it.

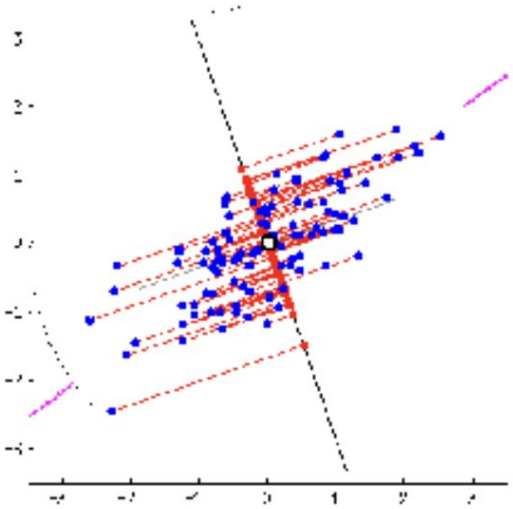
The specific example of PCA visualization

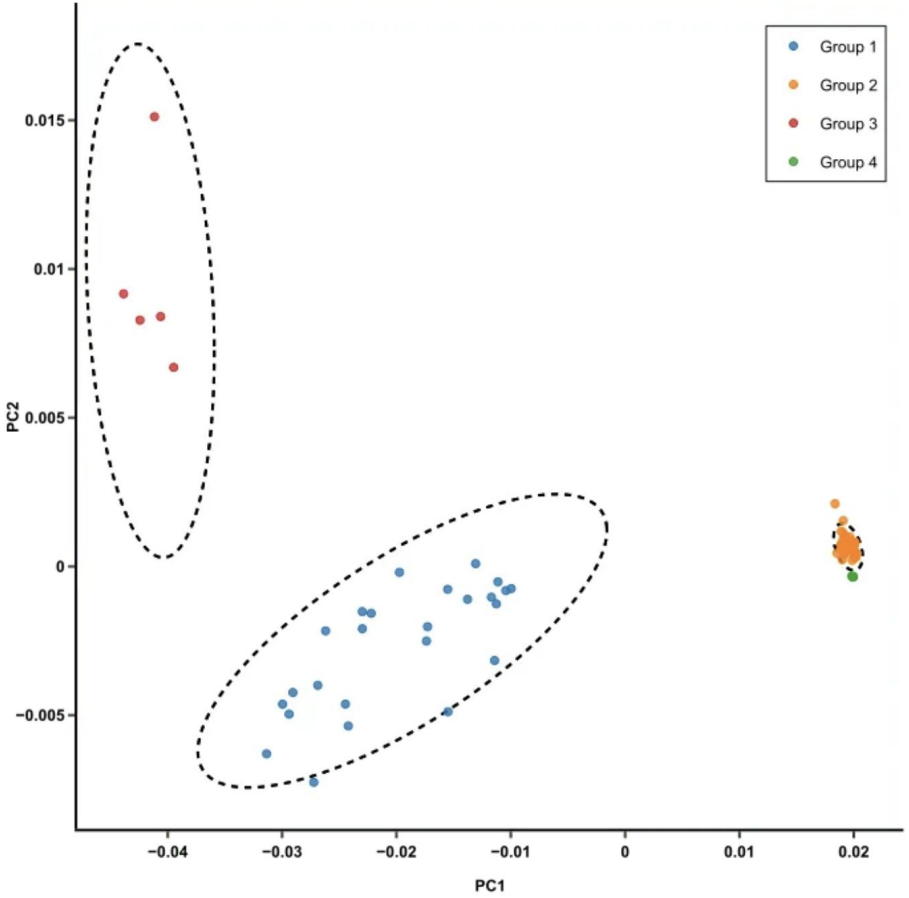
Background:

If we want to measure the expression of 15 genes from 60 mice, will will get a 15 x 60 table which is full of data. Additionally, each mouse is represented as a point in 15-dimensional space, which cannot visualize.

What PCA does:

We put a line in center of 15-D cloud of dots, and then it rotates in 15 directions on which the original 60 dots are projected to capture the maximum distance. The maximum distance is the most variation among 60 mice, and it fits to be PC1. And then we can find PC2.





Result:

After 15 genes of weight are determined, do the calculations to know where we put each mouse in PC1. Finally, PCA create a 2D scatter plot with PC1 and PC2 axes. Each dot represents each mouse with 15-gene profile. From this new plot, we can see that mice with similar expression patterns are clustered together, so there are 3 clusters, The new 2D plot reveals hidden patterns, making high-dimensional data easier to explore.

Thank You!

Work Cited

https://en.wikipedia.org/wiki/Principal_component_analysis

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

https://www.ibm.com/think/topics/principal-component-analysis

https://www.turing.com/kb/guide-to-principal-component-analysis

https://bioturing.medium.com/principal-component-analysis-explained-simply-894e8f6f4bfb

Process questions

Questions 1: We had a meeting before the presentation week to discuss our topics and then divide the parts. We met on Zoom.

Question 2: We didn’t assign anyone to be a manager. We divided the presentation into three parts, application, mathematics and visualization, according to the project rubric. Each one was responsible for one part. After finishing our own parts, we met again to help to check and modify each other’s content. Each member was satisfied with their parts and their workload.

Question 3: Even though we don’t understand PCA very clearly, which means it is a little bit hard for us, but we made our best efforts to contribute to this final project. Everyone is very modest and cooperative in the process of cooperation. We really enjoyed this final project.