Principal Component Analysis

PCA

🎯 1) Why do we need PCA?

Real-world datasets often have:

- Many features (high dimensionality),
- Redundant or correlated features,
- Noise.

This makes:

- Models slower,
- Patterns are harder to visualize,
- Risk of overfitting is higher.

PCA reduces dimensionality while **preserving as much variance (information) as possible**.

In simple words:

> PCA finds a new coordinate system (new axes) where the data spreads out the most and drops the least informative dimensions.

🧠 2) What does PCA do?

Given data with many features (x_1, x_2, ..., x_n):

- PCA **rotates the coordinate system** so that:
  - The **1st principal component** is the direction of **maximum variance**.
  - The **2nd PC** is the direction of maximum variance **perpendicular to the 1st**.
  - And so on.

Then we choose the **top k** principal components to reduce dimensions.

📚 3) Prerequisite Math (explained simply)

To understand PCA, we need 3 core concepts:

(A) **Variance**

Variance tells how spread out data is.

$Var(X) = \frac{1}{n}\sum(x_i - \bar{x})^2$

Higher variance = more information.

(B) **Covariance**

Covariance measures how two features change together.

$Cov(X, Y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$

- Positive → they increase together.
- Negative → one increases while the other decreases.
- Zero → unrelated.

This forms the **covariance matrix**:

$$\Sigma = \begin{bmatrix} Cov(x_1, x_1) & Cov(x_1, x_2) & \cdots \\ Cov(x_2, x_1) & Cov(x_2, x_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

(C) **Eigenvalues and Eigenvectors**

For a matrix A:

$Av = \lambda v$

- v = eigenvector → *direction*
- λ = eigenvalue → *strength (importance) of that direction*

In PCA:

- **Eigenvectors** = **principal directions (axes)**
- **Eigenvalues** = **amount of variance captured by each axis**

🔢 4) PCA Algorithm (Step-by-Step)

Input: Data matrix X with features.

1. Standardize the data
   (Center each feature to have mean 0).

2. Compute the covariance matrix Σ of X.

3. Compute eigenvalues and eigenvectors of Σ.

4. Sort eigenvectors by descending eigenvalues.
   (Higher eigenvalue = more variance captured.)

5. Select top k eigenvectors → W (projection matrix).

6. Transform data:
   X_reduced = X · W

Output: Data represented in fewer dimensions.

🎨 5) Visual Intuition

Imagine a cloud of points in 2D:

```
  *
 *  *
*   *
  *
```

- The longest stretch direction = **first principal component**.

- The second perpendicular direction = the **second principal component**.

If you **project** onto the first component, you effectively convert 2D → 1D **while keeping most information**.

🧠 Key Takeaway

Concept	Meaning in PCA
Variance	What PCA tries to preserve
Covariance	Shows relationships between features
Eigenvectors	New feature directions (principal components)
Eigenvalues	Importance/weight of each new direction
Dimensionality Reduction	Keep only the most informative components

---

🔥 Why PCA is Useful

Benefit	Explanation
Removes redundancy	Removes correlation between features
Speeds up ML models	Fewer features = faster training
Reduces noise	Small variance components can be dropped
Helps visualization	Convert 100+ features → 2 or 3 dimensions
Reduces overfitting	Less complexity → better generalization

✅ PCA (Step-by-Step Explanation Using Class Notation Only)

We are given data points:

$x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

---

1) Compute the mean of the data

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- This gives the dataset's center.

---

2) Center the data

For each data point:

$$\tilde{x}_i = x_i - \bar{\mu} \qquad (1 \leq i \leq n)$$

- This shifts the data so mean becomes zero.

---

3) Compute the covariance matrix

$$S = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T$$

Where:

- $S \in \mathbb{R}^{d \times d}$
- S is **symmetric and positive semi-definite**

This matrix tells us **how features vary together**.

---

4) Eigenvalue decomposition of the covariance matrix

$$S = V \Sigma V^T$$

Where:

- V contains **eigenvectors** of S
- \Sigma is a diagonal matrix containing **eigenvalues**

Write the eigenvalues in **descending order**:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Let the corresponding eigenvectors be:

$$\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_d$$

$$Each(\bar{u}_k \in \mathbb{R}^d).$$

---

5) Choose the top p eigenvectors

Pick the first p eigenvectors (those with the largest eigenvalues):

$$\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_p$$

These are the **principal components**.

---

6) Form the projection matrix

$$U = [, \bar{u}_1 \; \bar{u}_2 \; \ldots \; \bar{u}_p,]$$

- $U \in \mathbb{R}^{d \times p}$
- Each column is a principal direction.

---

7) Project the centered data onto the new subspace

$$U^T \tilde{x}_i \in \mathbb{R}^p$$

This is the **p-dimensional representation** of the original point x_i.

Or written explicitly:

$$U^T \tilde{x}_i = \begin{bmatrix} \bar{u}_1^T \tilde{x}_i \\ \bar{u}_2^T \tilde{x}_i \\ \vdots \\ \bar{u}_p^T \tilde{x}_i \end{bmatrix}$$

This is your **final reduced feature vector**.

🎯 Final Interpretation (in class language)

Object	Meaning
$\bar{\mu}$	Mean of data
$\tilde{x}_i$	Centered data
S	Covariance matrix
λ_k	Variance explained by principal direction $\bar{u}_k$
$\bar{u}_k$	Principal component direction
$U^T \tilde{x}_i$	Projection of x_i into lower dimension

🧠 Intuition (using the same symbols)

- The covariance matrix S tells us the directions in which the data spreads.
- Eigenvectors $\bar{u}_k$ of S point along those directions.
- Eigenvalues λ_k tell how *important* each direction is.
- We keep the **top p** directions (largest λ_k).
- We project data onto those directions → **dimension reduced**.

Example Numerical

We'll use four 2-D points and reduce to **p = 1** dimension:

$$x_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

1) Compute the mean

$$\bar{\mu} = \frac{1}{4} \sum_{i=1}^{4} \bar{x}_i = \begin{bmatrix} \frac{2+0+3+4}{4} \\ \frac{0+2+3+4}{4} \end{bmatrix} = \begin{bmatrix} 2.25 \\ 2.25 \end{bmatrix}$$

2) Center the data

$$\tilde{x}_i = \bar{x}_i - \bar{\mu}$$

$$\tilde{x}_1 = \begin{bmatrix} -0.25 \\ -2.25 \end{bmatrix},\ \tilde{x}_2 = \begin{bmatrix} -2.25 \\ -0.25 \end{bmatrix},\ \tilde{x}_3 = \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix},\ \tilde{x}_4 = \begin{bmatrix} 1.75 \\ 1.75 \end{bmatrix}$$

3) Covariance matrix (use your class convention $\frac{1}{n}$)

$$S = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T = \frac{1}{4}\begin{bmatrix} 8.75 & 4.75 \\ 4.75 & 8.75 \end{bmatrix} = \begin{bmatrix} 2.1875 & 1.1875 \\ 1.1875 & 2.1875 \end{bmatrix}$$

(S is symmetric and positive semidefinite.)

4) Eigenvalue decomposition

$S = V\Sigma V^T$

For this symmetric matrix,

$$\lambda_1 = 2.1875 + 1.1875 = 3.375, \quad \lambda_2 = 2.1875 - 1.1875 = 1.0$$

$$u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \bar{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

5) Pick top (p) eigenvectors (principal components)

With $(p = 1) : U = [\bar{u}_1]$.

Variance explained by PC1:

$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{3.375}{3.375+1.0} \approx 0.771 \quad (\approx 77.1\%)$

6) Projection matrix and projections

Projection to $(\mathbb{R}^p) : (z_i = U^T\tilde{x}_i = \bar{u}_1^T\tilde{x}_i)$.

Since ($\bar{u}_1 = \frac{1}{\sqrt{2}}[1, 1]^T$),

$z_i = \frac{1}{\sqrt{2}}(\tilde{x}_{i,1} + \tilde{x}_{i,2})$

Numerically:

- $z_1 = \frac{1}{\sqrt{2}}(-0.25 - 2.25) = -\frac{2.5}{\sqrt{2}} \approx -1.7678$
- $z_2 = \frac{1}{\sqrt{2}}(-2.25 - 0.25) = -1.7678$
- $z_3 = \frac{1}{\sqrt{2}}(0.75 + 0.75) = \frac{1.5}{\sqrt{2}} \approx 1.0607$
- $z_4 = \frac{1}{\sqrt{2}}(1.75 + 1.75) = \frac{3.5}{\sqrt{2}} \approx 2.4749$

So your 1-D representations are $z = [-1.7678,, -1.7678,, 1.0607,, 2.4749]^T$.

7) Final form (optional reconstruction)

Approximate each original point from the (p)-D code:

$\hat{x}_i = \bar{\mu} + U z_i = \bar{\mu} + \bar{u}_1 z_i$

✅ PCA as an Optimization Problem

We first consider **p = 1**, i.e., we want **one principal component** $\bar{u}_1$.

---

(1) Define the mean of the projected data

When data x_i is projected onto a direction $\bar{u}_1$,

The projection value is $\bar{u}_1^T x_i$.

So the mean of the projected values is:

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} (\bar{u}_1^T x_i)$$

This is simply the **mean along the projected axis**.

---

(2) Compute the variance of the projected data

Variance measures how spread out the projected data is:

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{u}_1^T x_i - \bar{u}_1^T \bar{\mu})^2$$

Your professor expands and simplifies this expression, eventually obtaining:

$$\boxed{\text{Var} = \bar{u}_1^T S \bar{u}_1}$$

where,

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mu})(x_i - \bar{\mu})^T$$

is the **covariance matrix**.

This is your key identity ((*)).

---

(3) Argue that S is symmetric

Let the centered data matrix be:

$$X = [, x_1 - \bar{\mu},\ x_2 - \bar{\mu},\ \ldots,\ x_n - \bar{\mu},]$$

Then:

$$S = \frac{1}{n} X X^T$$

Compute transpose:

$$S^T = \frac{1}{n}(X X^T)^T = \frac{1}{n} X X^T = S$$

So:

$$\boxed{S \text{ is symmetric}}$$

---

(4) Formulate PCA as an optimization problem

We want to **maximize the variance of the projection**:

$$\boxed{\text{maximize } \bar{u}_1^T S \bar{u}_1}$$

But we also require $|\bar{u}_1| = 1$, otherwise scaling ($\bar{u}_1$) increases variance artificially.

So:

$$\boxed{\max_{\bar{u}_1} \ \bar{u}_1^T S \bar{u}_1 \quad \text{s.t.} \quad \bar{u}_1^T \bar{u}_1 = 1}$$

This is **Problem P1** in your notes.

---

(5) Substitute eigenvalue decomposition

Since S is symmetric, we can write:

$$S = V \Sigma V^T$$

where:

- $V = [\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_d]$ are eigenvectors,
- $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_d)$ are eigenvalues.

Then:

$$\bar{u}_1^T S \bar{u}_1 = \bar{u}_1^T V \Sigma V^T \bar{u}_1$$

This creates **Problem P2** in your notes.

---

(6) Express $\bar{u}_1$ in eigenbasis

Because eigenvectors form an **orthonormal basis**, any vector ($\bar{u}_1$) can be written as:

$$\bar{u}_1 = \alpha_1 \bar{v}_1 + \alpha_2 \bar{v}_2 + \cdots + \alpha_d \bar{v}_d$$

or compactly:

$$\bar{u}_1 = V\alpha$$

with:

$$\sum_{i=1}^{d} \alpha_i^2 = 1$$

---

(7) Substitute into the optimization objective

$$\bar{u}_1^T S \bar{u}_1 = (V\alpha)^T V \Sigma V^T (V\alpha) = \alpha^T \Sigma \alpha$$

So P2 becomes P3:

$$\boxed{\max_{\alpha} \sum_{i=1}^{d} \alpha_i^2 \lambda_i \quad \text{s.t.} \quad \sum_{i=1}^{d} \alpha_i^2 = 1}$$

---

(8) Solve the optimization

Because ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$):

The quantity ($\sum \alpha_i^2 \lambda_i$) is largest when:

$\alpha_1 = 1, \quad \alpha_2 = \cdots = \alpha_d = 0$

---

(9) Therefore

$\boxed{\bar{u}_1 = \bar{v}_1}$

The **first principal component** is the **eigenvector of S** corresponding to the **largest eigenvalue**.

---

(10) Generalization to (p) dimensions

To find the next components:

- Maximize ($\bar{u}_2^T S \bar{u}_2$)
- Subject to:

- $\bar{u}_2^T \bar{u}_2 = 1$
- $\bar{u}_2^T \bar{u}_1 = 0$

This leads to:

$$\boxed{\bar{u}_2 = \bar{v}_2,\ \bar{u}_3 = \bar{v}_3, \ldots}$$

So PCA simply picks the **top p eigenvectors**.

🎉 Final Interpretation

> PCA finds the direction in which the projected data has the maximum variance.
>
> Mathematically, this is the eigenvector corresponding to the **largest eigenvalue** of the covariance matrix.
>
> Additional principal components are the remaining eigenvectors in decreasing order of eigenvalues, and are constrained to be orthogonal.
