CS 412 — Introduction to Machine Learning (UIC) April 03, 2025

Lecture 20

Instructor: Aadirupa Saha *Scribe(s): Simran Mishra*

Overview

In the last lecture, we covered the following main topics:

1. Introduction to PCA
2. Prelims of PCA
3. PCA Objective

This lecture focuses on:

1. Eigenvalues and Eigenvectors
2. PCA via Covariance Matrix and Eigenvectors
3. PCA on Symmetric Data Aligned with Diagonal Directions
4. PCA-Loss Minimization and Eigen Decomposition
5. PCA via Maximum Variance Formulation

1 Eigenvalues and Eigenvectors

1.1 What are Eigenvalues and Eigenvectors?

Eigenvectors and eigenvalues are fundamental concepts in linear algebra that describe how matrices transform vectors.

Theorem 1.1: Definition of Eigenvector and Eigenvalue

Given a square matrix $A \in \mathbb{R}^{n \times n}$, a non-zero vector $\vec{v} \in \mathbb{R}^n$ is called an eigenvector of A if:

$$A\vec{v} = \lambda\vec{v}$$

for some scalar $\lambda \in \mathbb{R}$. The scalar λ is called the eigenvalue corresponding to eigenvector $\vec{v}$.

This means that applying the matrix A to $\vec{v}$ does not change its direction—only its magnitude (scaled by λ).

1.2 Example: Diagonal Matrix

Let us consider the matrix:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

We want to find its eigenvalues and eigenvectors.

Exercise 1.1: Step 1: Find Eigenvalues

Solve the characteristic equation:

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 0 \\ 0 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda) = 0$$

Thus, the eigenvalues are:

$$\lambda_1 = 2, \quad \lambda_2 = 3$$

Exercise 1.2: Step 2: Find Eigenvectors

We solve $(A - \lambda I)\vec{v} = 0$ for each eigenvalue.
For $\lambda = 2$:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad 2I = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$A - 2I = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Now solve:

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow 0x + 1y = 0 \Rightarrow y = 0$$

So x can be any value. Choose $x = 1$, giving:

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

For $\lambda = 3$:

$$3I = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \quad A - 3I = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} - \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

Now solve:

$$\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow -1x = 0 \Rightarrow x = 0$$

So y can be any value. Choose $y = 1$, giving:

$$\vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

1.3 Conclusion

- $\lambda = 2$ has eigenvector $\vec{v}_1 = [1, 0]^T$
- $\lambda = 3$ has eigenvector $\vec{v}_2 = [0, 1]^T$

These vectors lie on the x-axis and y-axis respectively and only get scaled by A; they do not rotate. That's why they are eigenvectors—they retain their direction under the transformation.

Exercise 1.3: Practice Problem

Let

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$$

Find the eigenvalues and corresponding eigenvectors of A.

2 PCA via Covariance Matrix and Eigenvectors

2.1 2.1 Covariance Matrix Definition

In PCA, we begin by computing the covariance matrix S of the data:

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

Where:

- $x_i \in \mathbb{R}^d$ is the i^{th} data point
- $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the sample mean vector

2.2 2.2 Symmetric Distribution Example

Now consider a 2D dataset with points located only at $\pm e_1$ and $\pm e_2$, where:

- N_1: Number of points on $\pm e_1$
- N_2: Number of points on $\pm e_2$

By symmetry:

$$\bar{x} = \frac{1}{N}(N_1 e_1 + N_1(-e_1) + N_2 e_2 + N_2(-e_2)) = 0$$

So the covariance matrix simplifies to:

$$S = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$$

2.3 2.3 Compute Covariance Components

We now compute the individual outer products:

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow e_1 e_1^T = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow e_2 e_2^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Hence the full covariance matrix becomes:

$$S = \frac{1}{N} \left[2N_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 2N_2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right] = \frac{1}{N} \begin{bmatrix} 2N_1 & 0 \\ 0 & 2N_2 \end{bmatrix}$$

2.4 2.4 Eigenvalues and Eigenvectors of S

We now find the eigenvalues and eigenvectors by solving the characteristic equation of matrix S.
Let:

$$S = \frac{1}{N} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \quad \text{where } A = 2N_1, \quad B = 2N_2$$

Step 1: Characteristic Equation

We use the identity matrix $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and compute:

$$\det(S - \lambda I) = \det \left(\begin{bmatrix} \frac{A}{N} - \lambda & 0 \\ 0 & \frac{B}{N} - \lambda \end{bmatrix} \right) = \left(\frac{A}{N} - \lambda \right) \left(\frac{B}{N} - \lambda \right)$$

$$\Rightarrow \lambda_1 = \frac{A}{N} = \frac{2N_1}{N}, \quad \lambda_2 = \frac{B}{N} = \frac{2N_2}{N}$$

Step 2: Solve for Eigenvectors

We now solve $(S - \lambda I)\vec{v} = 0$ for each eigenvalue.

Eigenvector for λ_1:

$$S - \lambda_1 I = \begin{bmatrix} \frac{A}{N} - \frac{A}{N} & 0 \\ 0 & \frac{B}{N} - \frac{A}{N} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{B-A}{N} \end{bmatrix}$$

We solve:

$$\begin{bmatrix} 0 & 0 \\ 0 & \frac{B-A}{N} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \frac{B-A}{N} \cdot y = 0 \Rightarrow y = 0$$

x is free to choose. Let $x = 1$, then:

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Eigenvector for λ_2:

$$S - \lambda_2 I = \begin{bmatrix} \frac{A}{N} - \frac{B}{N} & 0 \\ 0 & \frac{B}{N} - \frac{B}{N} \end{bmatrix} = \begin{bmatrix} \frac{A-B}{N} & 0 \\ 0 & 0 \end{bmatrix}$$

We solve:

$$\begin{bmatrix} \frac{A-B}{N} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \frac{A-B}{N} \cdot x = 0 \Rightarrow x = 0$$

Choose $y = 1$, then:

$$\vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Summary

- $\lambda_1 = \frac{2N_1}{N}$ with eigenvector $\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
- $\lambda_2 = \frac{2N_2}{N}$ with eigenvector $\vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Theorem 2.1: Principal Component Selection

The direction that captures the most variance is the eigenvector corresponding to the largest eigenvalue. If $N_1 \gg N_2$, then $\lambda_1 \gg \lambda_2$, so $\vec{v}_1$ is the first principal component.

2.5 2.5 PCA Intuition Recap

- Covariance matrix captures how features vary together.
- Its eigenvectors represent directions of maximum and minimum variance.
- Eigenvalues tell how much variance exists in each direction.
- PCA picks top m eigenvectors with highest eigenvalues for projection.

Exercise 2.1: Eigenvectors of a Diagonal Covariance Matrix

Let:

$$S = \frac{1}{6} \begin{bmatrix} 12 & 0 \\ 0 & 6 \end{bmatrix}$$

1. Compute the eigenvalues using the characteristic equation.
2. Solve $(S - \lambda I)\vec{v} = 0$ to get eigenvectors.
3. Which principal component direction will PCA select and why?

3 PCA on Symmetric Data Aligned with Diagonal Directions

3.1 Covariance Matrix with Diagonal Directions

Assume we have a dataset where the data points lie along the directions $v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, possibly negated. We denote:

- N_1: number of points along v_1 and $-v_1$
- N_2: number of points along v_2 and $-v_2$
- $N = N_1 + N_2$: total number of data points

The general covariance formula is:

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

Since the dataset is symmetric about the origin (equal number of points in opposite directions), the mean is:

$$\bar{x} = \frac{1}{N} \sum x_i = 0 \quad \Rightarrow \quad S = \frac{1}{N} \sum x_i x_i^T$$

Now, normalize the vectors to unit length:

$$\hat{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \hat{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Step: Compute Outer Products (Rank-1 Matrices)

$$\hat{v}_1 \hat{v}_1^T = \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\hat{v}_2 \hat{v}_2^T = \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Covariance Matrix: Add Contributions

$$S = \frac{1}{N} \left(N_1 \cdot \hat{v}_1 \hat{v}_1^T + N_2 \cdot \hat{v}_2 \hat{v}_2^T \right) = \frac{1}{N} \left(\frac{N_1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{N_2}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right)$$

$$= \frac{1}{N} \begin{bmatrix} 2(N_1 + N_2) & 2(N_1 - N_2) \\ 2(N_1 - N_2) & 2(N_1 + N_2) \end{bmatrix}$$

Let $A = 2(N_1 + N_2), \quad B = 2(N_1 - N_2)$, so:

$$S = \frac{1}{N} \begin{bmatrix} A & B \\ B & A \end{bmatrix}$$

3.2 Eigenvalue Computation (Step-by-step)

We solve the characteristic equation:

$$\det(S - \lambda I) = 0 \Rightarrow \det\left(\begin{bmatrix} \frac{A}{N} - \lambda & \frac{B}{N} \\ \frac{B}{N} & \frac{A}{N} - \lambda \end{bmatrix}\right) = 0$$

Compute determinant:

$$\left(\frac{A}{N} - \lambda\right)^2 - \left(\frac{B}{N}\right)^2 = 0 \Rightarrow \left(\frac{A}{N} - \lambda\right)^2 = \left(\frac{B}{N}\right)^2 \Rightarrow \lambda = \frac{A \pm B}{N}$$

Substitute back:

$$\lambda_1 = \frac{2(N_1 + N_2) + 2(N_1 - N_2)}{N} = \frac{4N_1}{N}$$

$$\lambda_2 = \frac{2(N_1 + N_2) - 2(N_1 - N_2)}{N} = \frac{4N_2}{N}$$

3.3 Eigenvector Computation

Eigenvector for $\lambda_1 = \frac{4N_1}{N}$

$$S - \lambda_1 I = \frac{1}{N}\begin{bmatrix} A - (A + B) & B \\ B & A - (A + B) \end{bmatrix} = \frac{1}{N}\begin{bmatrix} -B & B \\ B & -B \end{bmatrix}$$

Solve:

$$\begin{bmatrix} -B & B \\ B & -B \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow -Bx + By = 0 \Rightarrow x = y$$

Choose $x = 1, y = 1 \Rightarrow \vec{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Eigenvector for $\lambda_2 = \frac{4N_2}{N}$

$$S - \lambda_2 I = \frac{1}{N}\begin{bmatrix} A - (A - B) & B \\ B & A - (A - B) \end{bmatrix} = \frac{1}{N}\begin{bmatrix} B & B \\ B & B \end{bmatrix}$$

Solve:

$$\begin{bmatrix} B & B \\ B & B \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow Bx + By = 0 \Rightarrow x = -y$$

Choose $x = -1, y = 1 \Rightarrow \vec{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$

3.4 Interpretation

- If $N_1 > N_2$, then $\lambda_1 > \lambda_2$
- So PCA selects $\vec{v}_1$ (diagonal direction [1, 1]) as the first principal component.
- The directions are orthogonal and unit norm, ideal for PCA basis.

4 PCA — Loss Minimization and Eigen Decomposition

4.1 Motivation: Minimizing Reconstruction Error

Let the dataset be $\{x_n\}_{n=1}^N \subset \mathbb{R}^D$. Our goal is to find projections $\tilde{x}_n \in \mathbb{R}^D$ that minimize the reconstruction loss:

$$J = \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|_2^2$$

We write the reconstruction using m principal directions:

$$\tilde{x}_n = \sum_{i=1}^{m} z_{ni} u_i + \sum_{j=m+1}^{D} b_j u_j$$

where:

- $z_{ni} = \langle x_n, u_i \rangle$
- $b_j = \langle \bar{x}, u_j \rangle$

4.2 Reformulation Using Covariance Matrix

Given that $\bar{x} = \frac{1}{N} \sum x_n$, and $S = \frac{1}{N} \sum (x_n - \bar{x})(x_n - \bar{x})^\top$, the error becomes:

$$J = \sum_{j=m+1}^{D} u_j^\top S u_j$$

We aim to minimize this quantity by selecting the appropriate u_j's.

4.3 Recursive Derivation — Case 1: $m = D - 1$

Let u_D be the vector that minimizes $u_D^\top S u_D$ subject to $\|u_D\| = 1$. This is a constrained optimization problem.
Using Lagrangian:

$$\mathcal{L}(u_D, \lambda) = u_D^\top S u_D + \lambda(1 - u_D^\top u_D)$$

Taking derivative and setting it to zero:

$$\nabla_{u_D} \mathcal{L} = 2 S u_D - 2\lambda u_D = 0 \Rightarrow S u_D = \lambda u_D$$

Hence, u_D must be an eigenvector of S, and λ is the corresponding eigenvalue. To minimize J, we choose u_D as the eigenvector corresponding to the **smallest** eigenvalue λ_D.
Then:

$$J = \lambda_D$$

4.4 Case 2: Recursive Selection of All u_i

Now assume $m = D - 2$. We have already selected u_D. We now find u_{D-1} such that:

$$\min_{u_{D-1}} u_{D-1}^\top S u_{D-1}, \quad \text{subject to:}$$

- $\|u_{D-1}\|^2 = 1$
- $u_{D-1}^\top u_D = 0$

This ensures u_{D-1} lies in the orthogonal subspace to u_D. The same Lagrangian trick shows:

$$S u_{D-1} = \lambda_{D-1} u_{D-1}$$

So u_{D-1} must be the eigenvector of S corresponding to the second smallest eigenvalue.
We continue this recursively:

$$u_i \text{ is the eigenvector of } S \text{ corresponding to the } i\text{-th smallest eigenvalue}$$

4.5 How to Choose m

We want to discard directions $u_{m+1}, \ldots, u_D$, and retain $u_1, \ldots, u_m$. The reconstruction error is:

$$J = \sum_{j=m+1}^{D} \lambda_j$$

Heuristics for choosing m:

- Find the smallest m such that:

  $$\lambda_m - \lambda_{m+1} < \varepsilon$$

  for some small threshold $\varepsilon > 0$
- Use the **elbow method**: plot $J_m = \sum_{j=m+1}^{D} \lambda_j$ and pick the point where the decrease flattens

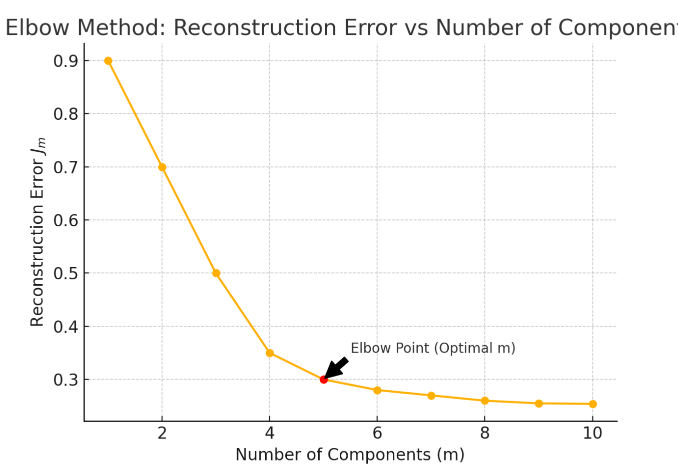

Figure 1: The elbow point corresponds to the optimal number of retained principal components.

4.6 Final PCA Summary

- PCA finds orthonormal directions $u_1, \dots, u_D$ which are eigenvectors of S
- Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ measure variance along each u_i
- To reduce dimension to m, keep $u_1, \dots, u_m$
- This minimizes reconstruction loss:

$$J = \sum_{j=m+1}^{D} \lambda_j$$

- Variance preserved is $\sum_{i=1}^{m} \lambda_i$

5 PCA via Maximum Variance Formulation

PCA can also be derived by finding the directions that **maximize variance** in the data, which turns out to be equivalent to minimizing reconstruction error.

5.1 Step 1: Projecting on a Direction

Let $\vec{v}_1 \in \mathbb{R}^D$ be a unit vector (i.e., $\|\vec{v}_1\| = 1$).
We project each centered data point $\bar{x}_n = x_n - \mu$ onto $\vec{v}_1$. The projection is a scalar:

$$z_{n1} = \vec{v}_1^\top \bar{x}_n$$

The goal is to find the direction $\vec{v}_1$ along which the **variance of projections** z_{n1} is maximized.

5.2 Step 2: Variance Maximization Objective

Define variance along $\vec{v}_1$:

$$\text{Var}(z_{n1}) = \frac{1}{N} \sum_{n=1}^{N} (\vec{v}_1^\top \bar{x}_n)^2 = \vec{v}_1^\top S \vec{v}_1$$

Subject to:

$$\|\vec{v}_1\|^2 = 1$$

This is a constrained optimization problem. Use Lagrange multipliers:

$$\mathcal{L}(\vec{v}_1, \lambda) = \vec{v}_1^\top S \vec{v}_1 - \lambda(\vec{v}_1^\top \vec{v}_1 - 1)$$

Take the derivative and set it to zero:

$$\nabla_{\vec{v}_1} \mathcal{L} = 2S\vec{v}_1 - 2\lambda \vec{v}_1 = 0 \quad \Rightarrow \quad S\vec{v}_1 = \lambda \vec{v}_1$$

So, $\vec{v}_1$ must be an eigenvector of S, and λ is its eigenvalue.

5.3 Step 3: Recursive Projection Directions

To find the second direction $\vec{v}_2$, maximize variance along it while keeping it orthogonal to $\vec{v}_1$:

$$\text{maximize } \vec{v}_2^\top S \vec{v}_2 \quad \text{subject to } \|\vec{v}_2\| = 1, \vec{v}_1^\top \vec{v}_2 = 0$$

This again leads to:

$$S\vec{v}_2 = \lambda_2 \vec{v}_2$$

Repeat this process recursively — so PCA directions $\vec{v}_1, \ldots, \vec{v}_m$ are the top m eigenvectors of S.

5.4 Equivalence to Reconstruction Error Minimization

The reconstruction error in PCA is:

$$J = \sum_{j=m+1}^{D} \lambda_j$$

The total variance of data is:

$$\text{Total Variance} = \sum_{j=1}^{D} \lambda_j$$

Variance retained using top m components:

$$\text{Retained Variance} = \sum_{j=1}^{m} \lambda_j$$

So minimizing the reconstruction error $\sum_{j=m+1}^{D} \lambda_j$ is equivalent to maximizing the retained variance $\sum_{j=1}^{m} \lambda_j$.

5.5 Geometric Interpretation

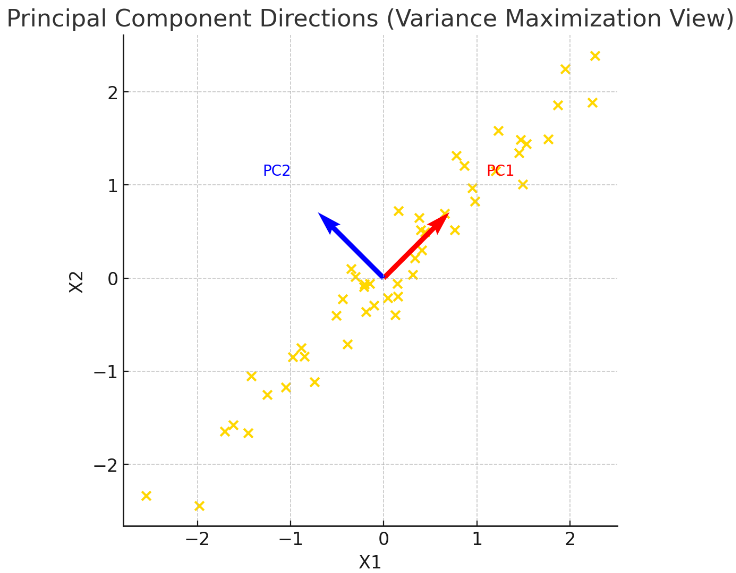

Figure 2: Principal component directions via variance maximization. PC1 (red) captures the highest variance; PC2 (blue) is orthogonal and captures remaining variance.

5.6 Final Remark

Conclusion

- PCA can be derived either by **minimizing reconstruction error** or by **maximizing projected variance**.
- Both formulations lead to the same eigenvalue problem.
- The top m eigenvectors of S provide an optimal subspace for projection.

Next Lecture

In the next lecture, we will cover the following main topics:

1. Introduction to Clustering
2. K Means Clustering
3. Spectral Clustering

References:

1. Chapter 12, Principal Component Analysis, from the book *Pattern Recognition and Machine Learning* by Christopher M. Bishop.

2. Lecture notes by Prof. Gilbert Strang from the course *Linear Algebra (18.06)* MIT OpenCourseWare Source