Q1. What is a projection and how is it used in PCA?

A **projection** is the process of mapping data points from a higher-dimensional space onto a lower-dimensional space along certain directions.

- In PCA, projection is used to **transform original features into new axes (principal components)** that capture the most variance.

- Mathematically: if x is a data point and w is a principal component vector, the projection is:

z=wTxz = w^T xz=wTx

where z is the coordinate of x along w.

Purpose in PCA: Reduce dimensionality while preserving the most important structure (variance) of the data.

Q2. How does the optimization problem in PCA work, and what is it trying to achieve?

- PCA solves an **optimization problem**: find a direction (principal component) that **maximizes the variance** of the projected data.

- Formally:

maximize Var(wTX)subject to ∣∣w∣∣=1\text{maximize } \text{Var}(w^T X) \quad \text{subject to } ||w|| = 1maximize Var(wTX)subject to ∣∣w∣∣=1

- This ensures the principal component captures the **maximum spread/variation** in the data.

- Subsequent components are chosen to be **orthogonal** to previous ones and capture the next largest variance.

Goal: Reduce dimensionality while retaining as much information (variance) as possible.

Q3. What is the relationship between covariance matrices and PCA?

- PCA uses the **covariance matrix** of the data to find directions of maximum variance.
- Covariance matrix Σ\SigmaΣ captures how features vary together.
- **Eigenvectors** of Σ\SigmaΣ = principal component directions
- **Eigenvalues** of Σ\SigmaΣ = variance along each principal component

Covariance matrix → eigen decomposition → principal components.

Q4. How does the choice of number of principal components impact the performance of PCA?

- **Too few components:** Lose important information → **underfitting**
- **Too many components:** Minimal dimensionality reduction → computational cost remains high, risk of **overfitting** if noise is included
- **Common approach:** Retain components that explain, e.g., **95–99% of variance**.

Q5. How can PCA be used in feature selection, and what are the benefits?

- PCA can **reduce the number of features** by projecting original features onto the top principal components.
  Benefits:

1. Reduces dimensionality → lower computational cost
2. Removes correlated/redundant features → improves model efficiency
3. Mitigates the **curse of dimensionality**
4. Reduces noise → can improve generalization

Note: PCA creates **new features (linear combinations)**, not a subset of original features.

Q6. Common applications of PCA in data science and machine learning

1. **Dimensionality reduction** for large datasets
2. **Visualization** of high-dimensional data in 2D or 3D
3. **Noise reduction / data denoising**
4. **Preprocessing before clustering, regression, or classification**
5. **Feature extraction** for machine learning models
6. **Image compression** (e.g., reducing pixels while retaining structure)

Q7. What is the relationship between spread and variance in PCA?

- **Spread of data** refers to how widely data points are distributed in a particular direction.
- **Variance** quantifies this spread numerically.
- PCA identifies directions where the **spread/variance is maximal** → these directions become principal components.

Maximum spread = maximum variance → first principal component.

Q8. How does PCA use the spread and variance of the data to identify principal components?

- PCA calculates the **covariance matrix** → measures spread/variance along all feature axes.
- Eigenvectors of covariance matrix = directions of maximum variance (spread)
- Eigenvalues = amount of variance captured in each direction
- Principal components are ordered from **highest variance (most spread)** to lowest

Idea: Projecting onto high-variance directions preserves most information.

Q9. How does PCA handle data with high variance in some dimensions but low variance in others?

- PCA automatically prioritizes **dimensions with higher variance**:
  - High variance dimensions contribute more to the first principal components.
  - Low variance dimensions contribute to later components or may be effectively ignored.
- This ensures PCA captures **most informative patterns** while reducing the impact of noisy or uninformative low-variance dimensions.

> Example: In a dataset with one feature ranging 0–1000 and another 0–1, PCA will favor the first unless data is scaled.

Important: Scaling features before PCA (standardization) is often necessary to prevent dominance of high-magnitude features.
