Eigenvectors and Eigenvalues: A Simple Example

Recall the fundamental equation for eigenvectors (v) and eigenvalues (λ) of a square matrix A: **Av = λv**

This means that when matrix A multiplies its eigenvector v, the result is simply the same eigenvector scaled by its corresponding eigenvalue λ. The direction of the eigenvector remains unchanged.

Let's consider a simple 2x2 square matrix A:

A = [ [2, 0], [0, 5] ]

For this matrix A, the eigenvectors and corresponding eigenvalues are:

* v₁ = [1, 0] with eigenvalue λ₁ = 2
  + Check: A \* v₁ = [[2, 0], [0, 5]] \* [1, 0] = [2\*1 + 0\*0, 0\*1 + 5\*0] = [2, 0] = 2 \* [1, 0] = λ₁ \* v₁
* v₂ = [0, 1] with eigenvalue λ₂ = 5
  + Check: A \* v₂ = [[2, 0], [0, 5]] \* [0, 1] = [2\*0 + 0\*1, 0\*0 + 5\*1] = [0, 5] = 5 \* [0, 1] = λ₂ \* v₂

So, this matrix A has two sets of eigenvectors/eigenvalues.

Eigenvectors of Covariance Matrix = Principal Components

Now, let's connect this back to PCA.

* **Goal Revisited:** Our goal in PCA is to find a new set of basis vectors (a new coordinate system) for our data such that when the data is represented in this new basis, its covariance matrix becomes **diagonal**. A diagonal covariance matrix means the new features (the principal components) are **uncorrelated** with each other.
* **The Key Insight:** It turns out that the **eigenvectors of the original data's Covariance Matrix (C)** provide exactly this new set of basis vectors!
* **Principal Components (PCs):** These eigenvectors of the covariance matrix **are** the Principal Components of the dataset.
  + v₁ (eigenvector corresponding to the largest eigenvalue λ₁) is **Principal Component 1 (PC1)**.
  + v₂ (eigenvector corresponding to the second-largest eigenvalue λ₂) is **Principal Component 2 (PC2)**, and so on.
* **Direction of Max Variance:** These eigenvectors represent the directions in the original feature space along which the data exhibits the most variance.

Ordering PCs and Selecting Components (Feature Vector)

The eigen-decomposition of the covariance matrix C yields the eigenvectors v₁, v₂, v₃, ... and their corresponding eigenvalues λ₁, λ₂, λ₃, ....

* **Variance Explained:** As noted earlier, the eigenvalues (λᵢ) quantify the **amount of variance** explained by their corresponding eigenvectors (vᵢ, the Principal Components).
* **Ordering:** We order the eigenvectors (Principal Components) based on their eigenvalues, from highest to lowest: λ₁ ≥ λ₂ ≥ λ₃ ≥ ... This means PC1 (v₁) captures the most variance, PC2 (v₂) captures the second most (orthogonal to PC1), and so forth.
* **Explained Variance Ratio:** Often, instead of looking at the raw eigenvalue, we look at the **percentage of variance explained** by each PC: Explained Variance Ratio for PCᵢ = λᵢ / (Σ λⱼ) (where the sum is over all eigenvalues).
* **Choosing How Many PCs to Keep:** By examining the eigenvalues or the explained variance ratios, we decide how many of the top principal components to retain to capture a sufficient amount of the total variance in the original dataset (e.g., choose enough PCs to explain 90%, 95%, or 99% of the variance). The chosen eigenvectors form the **Feature Vector**.

Step 4 & 5: Represent the Data using Chosen PCs (Transformation)

The final step of PCA is to use the chosen principal components (the selected eigenvectors that form the Feature Vector) to transform the original dataset into a lower-dimensional space.

* **Process:** This involves projecting the original (standardized) data onto the new axes defined by the selected principal component eigenvectors. Mathematically, this is typically done by taking the dot product of the original standardized data matrix and the matrix formed by the chosen eigenvectors (the Feature Vector).
* **Result:** This transformation **recasts the original dataset** using the new Principal Component axes. If we started with N features and decided to keep k principal components (k < N), the resulting transformed dataset will have only k dimensions (features).
* **Example:** If the original dataset had 70 dimensions (features) and we found that the top 6 Principal Components explain 95% of the variance, we might choose to keep only those top 6 PCs. The transformed dataset would then have only 6 features, while retaining 95% of the information (variance) from the original 70 features.
* **Usage:** This lower-dimensional, transformed dataset can then be used for visualization, further machine learning model training (potentially leading to faster training and reduced overfitting), or other analyses.

PCA Process Revisited (Flowchart Summary)

1. **Original Dataset** (M x N, M samples, N features)
2. **(Standardize Data)** - *Implicitly needed before Covariance*
3. **Find Covariance Matrix** (N x N)
4. **Find Eigenvectors (PCs) and Eigenvalues** (N eigenvectors, N eigenvalues)
5. **Rank PCs** according to High to Low values of Eigenvalues
6. **(Select Top k PCs)** - Based on explained variance
7. **Recast Original Dataset** using chosen top k PCs -> **Transformed Dataset** (M x k)

Explained Variance Ratio & Ranking PCs (Scree Plot)

* **Recap:** Eigenvalues indicate the variance explained by each principal component (eigenvector). Higher eigenvalue = more variance explained = more important direction.
* **Ranking:** Listing the Principal Components based on the higher-to-lower values of their corresponding eigenvalues gives us an ordered list reflecting their significance.
* **Decision Point:** The crucial decision is how many of these top components to retain to capture enough information while reducing dimensionality.
* **Scree Plot:** A **Scree Plot** is a common visualization used to help make this decision. It plots:
  + The eigenvalue (or percentage of variance explained) for each principal component (usually on the y-axis).
  + The principal components, ordered from 1 to N (on the x-axis).
* **Interpretation:**
  + The plot typically shows a steep drop-off initially (the first few PCs capture a large amount of variance) followed by a leveling off (later PCs capture progressively less variance).
  + Similar to the Elbow Method in clustering, we look for the "elbow" point in the scree plot – the point where adding another principal component results in only a marginal increase in the total explained variance. The number of components at or just before the elbow is often chosen as the optimal number (k) to retain.
  + Alternatively, one can calculate the *cumulative* explained variance and choose the number of components (k) required to exceed a certain threshold (e.g., 90%, 95%). The cumulative curve is often overlaid on the scree plot (as seen in the example).

By analyzing the explained variance ratios and the scree plot, we can make an informed decision about the trade-off between dimensionality reduction and information retention.