CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Eigendecomposition and PCA

Date: October 15th, 2025 *Author: Surbhi Goel*

1 Recap and Preview

Two lectures ago we saw that the same linear transformation can look very different depending on which basis we use. In particular, the motivating example showed $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ becoming diagonal in a different basis. Last time we added geometric structure through inner products and projections. Today: *which linear transformations can be diagonalized, and how do we find the right basis?*

2 Eigenvalues and Eigenvectors

2.1 General Definition

- Consider a linear transformation $\Phi : V \to V$ (an **endomorphism**) on a vector space V. A scalar $\lambda \in \mathbb{R}$ is an **eigenvalue** and a non-zero vector $v \in V$ is a corresponding **eigenvector** if:
$$\Phi(v) = \lambda v$$
- **Intuition:** Most vectors change direction when transformed. But eigenvectors are special—they only get scaled by λ, without changing direction.
- This means eigenvalues and eigenvectors identify the directions along which the transformation is particularly simple (just scaling).

2.2 Matrix Representation

For finite-dimensional V with basis B, let A be the transformation matrix of Φ with respect to B:

- If $v \in V$ is an eigenvector with eigenvalue λ, then $\Phi(v) = \lambda v$. In coordinate form, this becomes $A\hat{v} = \lambda\hat{v}$ where $\hat{v} \in \mathbb{R}^n$ is the coordinate vector of v with respect to B.
- **Key observation:** If matrices A and $\tilde{A}$ represent the same transformation in different bases (related by $\tilde{A} = S^{-1}AS$), they have the same eigenvalues. Eigenvalues are properties of the *transformation*, not the choice of basis or matrix representation.
- If v is an eigenvector, so is any scalar multiple cv (for $c \neq 0$). We typically normalize eigenvectors to have unit length.

2.3 Computing Eigenvalues and Properties

For an $n \times n$ matrix A:

- From $A\hat{v} = \lambda\hat{v}$, we get $(A - \lambda I)\hat{v} = 0$. For non-trivial solutions, $A - \lambda I$ must be singular, giving the **characteristic equation**: $\det(A - \lambda I) = 0$.
- An $n \times n$ matrix has at most n eigenvalues (counting multiplicities).
- A and A^T have the same eigenvalues (though possibly different eigenvectors).
- For symmetric matrices ($A = A^T$): all eigenvalues are real. If additionally positive (semi-)definite, all eigenvalues are positive (non-negative).
- **Useful facts:** $\sum_{i=1}^{n} \lambda_i = \text{tr}(A)$ and $\prod_{i=1}^{n} \lambda_i = \det(A)$.

3 Eigendecomposition

3.1 Diagonalization

Recall from Lecture 12: when we change basis via S, the transformation matrix changes as $D = S^{-1}AS$. The question is: can we choose S to make D diagonal?

- A linear transformation $\Phi : V \to V$ (or its matrix representation A) is **diagonalizable** if there exists a basis of eigenvectors.
- In matrix form, this means we can write:

  $$A = PDP^{-1} \quad \text{where} \quad D = \text{diag}(\lambda_1, \ldots, \lambda_n)$$

  with eigenvalues λ_i on the diagonal and corresponding eigenvector coordinate vectors $\hat{v}_i$ as columns of P.
- **Why?** Write $P = [\hat{v}_1 | \cdots | \hat{v}_n]$ where $A\hat{v}_i = \lambda_i \hat{v}_i$. Then:

  $$AP = [A\hat{v}_1 | \cdots | A\hat{v}_n] = [\lambda_1 \hat{v}_1 | \cdots | \lambda_n \hat{v}_n] = PD$$

  Therefore $A = PDP^{-1}$. The matrix P changes from the eigenbasis (where the transformation is diagonal) to the standard basis.
- A transformation is diagonalizable if and only if it has n linearly independent eigenvectors.

3.2 Spectral Theorem

Not every transformation is diagonalizable. But when we add inner product structure, symmetric transformations have a special property.

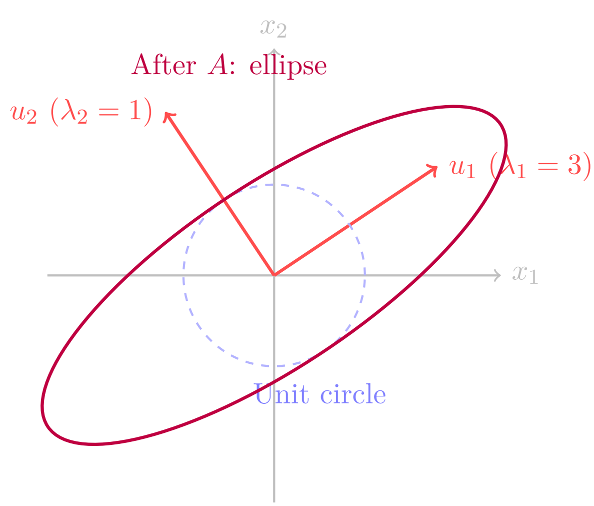
Theorem 1 (Spectral Theorem)**.** *If $\Phi : V \to V$ is a linear transformation on an inner product space with Φ symmetric (i.e., $\langle \Phi(u), v\rangle = \langle u, \Phi(v)\rangle$ for all $u, v \in V$), then there exists an orthonormal basis of eigenvectors.*

For matrices: if $A \in \mathbb{R}^{n\times n}$ is symmetric, then $A = Q\Lambda Q^T$ where Q is orthogonal (columns are orthonormal eigenvectors) and $\Lambda = diag(\lambda_1, \dots, \lambda_n)$ contains the real eigenvalues.

- This strengthens diagonalizability: symmetric transformations have an *orthonormal* eigenbasis, so $P^{-1} = Q^T$.
- **Why orthonormal?** The key insight is that eigenvectors for distinct eigenvalues are automatically orthogonal:
  - If $A\hat{v}_1 = \lambda_1\hat{v}_1$ and $A\hat{v}_2 = \lambda_2\hat{v}_2$ with $\lambda_1 \neq \lambda_2$, then:

    $$\lambda_1\hat{v}_1^T\hat{v}_2 = (A\hat{v}_1)^T\hat{v}_2 = \hat{v}_1^T A^T\hat{v}_2 = \hat{v}_1^T A\hat{v}_2 = \hat{v}_1^T(\lambda_2\hat{v}_2) = \lambda_2\hat{v}_1^T\hat{v}_2$$

    Since $\lambda_1 \neq \lambda_2$, we must have $\hat{v}_1^T\hat{v}_2 = 0$.
  - For repeated eigenvalues, the eigenspace can have dimension > 1, but we can always choose an orthogonal basis within it.
  - Finally, normalize all eigenvectors to unit length to get an orthonormal basis.
- **Geometric interpretation:** $A = Q\Lambda Q^T$ means: (1) Q^T rotates to the orthonormal eigenbasis, (2) Λ scales along each eigendirection, (3) Q rotates back. Example: suppose $\lambda_1 = 3, \lambda_2 = 1$.



  The transformation A turns circles into ellipses aligned with the eigenvectors, stretching by the eigenvalues.

- Covariance matrices are symmetric and positive semi-definite, so they always have this decomposition with $\lambda_i \geq 0$.

3.3 Other Decompositions as Basis Change

Eigendecomposition is one of several powerful techniques that finds a new basis to simplify a linear transformation. Other decompositions offer different perspectives by changing basis in specific ways:

- **Singular Value Decomposition (SVD, $A = U\Sigma V^T$):** For any rectangular matrix, this finds two different orthonormal bases (for the domain and codomain, with respect to the standard inner product) that make the transformation diagonal. This is key to PCA's practical implementation.
- **QR Decomposition ($A = QR$):** For any rectangular matrix, this finds an orthonormal basis for the column space (Q) such that the transformation in this new basis is upper-triangular (R). This is useful for efficiently solving linear systems.
- **Cholesky Decomposition ($A = LL^T$):** For symmetric, positive-definite matrices, this can be viewed as finding a new basis that is orthonormal with respect to the inner product defined by A (i.e., $\langle x, y\rangle_A = x^T A y$). In this new basis, the transformation represented by A becomes the identity.

4 Application: Principal Component Analysis (PCA)

PCA finds a low-dimensional representation of high-dimensional data ($x_1, \ldots, x_N \in \mathbb{R}^D$) by exploiting the spectral theorem.

Setup: Center the data and compute its covariance matrix $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$. Since Σ is symmetric and positive semi-definite, the spectral theorem gives $\Sigma = Q\Lambda Q^T$ where:

- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_D)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$
- Columns of Q are orthonormal eigenvectors (the **principal components**)
- In the eigenbasis, the covariance is *diagonal*: $Q^T \Sigma Q = \Lambda$
- The diagonal entries λ_i are the variances along each eigendirection

The key idea: By changing to the eigenbasis (via Q^T), we transform correlated data into uncorrelated components. The eigenvectors with largest eigenvalues capture the most variance (information). To reduce dimensionality, project onto the top k eigenvectors: $z_i = U_k^T(x_i - \mu)$ where U_k contains the first k columns of Q.

4.1 Why Eigenvectors Maximize Variance

The goal is to find a direction u (unit vector) that maximizes the variance of the data when projected onto that direction. The projection of a data point $\tilde{x}_i$ onto u is $\tilde{x}_i^T u$, and the variance of these projections is:

$$\text{Var}(u) = \frac{1}{N}\sum_{i=1}^{N}(\tilde{x}_i^T u)^2 = u^T\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\tilde{x}_i^T\right)u = u^T\Sigma u$$

- Using the spectral decomposition $\Sigma = Q\Lambda Q^T$ (with $\lambda_1 \geq \cdots \geq \lambda_D$), we can rewrite the variance. Let $v = Q^T u$. Since Q is orthogonal, $\|v\| = 1$. The variance becomes:

$$u^T \Sigma u = (Qv)^T \Sigma (Qv) = v^T Q^T \Sigma Q v = v^T \Lambda v = \sum_{i=1}^{D} \lambda_i v_i^2$$

- Since $\sum v_i^2 = 1$, this expression is a convex combination of the eigenvalues and is maximized at λ_1 when $v = (1, 0, \ldots, 0)$.

- This corresponds to $u = Qv = u_1$, the first eigenvector of Σ. The second principal component, u_2, is found by maximizing the same variance objective within the orthogonal complement of $\text{span}(u_1)$. This process continues, with each subsequent component maximizing variance in the subspace orthogonal to all previously found components.

Next lecture: we extend these ideas to infinite-dimensional function spaces, leading to the Representer Theorem.