10. Singular Values, SVD and PCA

We begin with the fundamental concept of singular values and their associated decomposition.

Definition 1 (Singular Values)

For an $m \times n$ matrix A, the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$ are the square roots of the eigenvalues of $A^\top A$.

The singular values characterize important properties of linear transformations that eigenvalues alone cannot capture, especially for non-square matrices.

Definition 2 (Singular Value Decomposition (SVD))

Every $m \times n$ matrix A can be factorized as $A = U\Sigma V^\top$ where U is an $m \times m$ orthogonal matrix whose columns are the left singular vectors, V is an $n \times n$ orthogonal matrix whose columns are the right singular vectors, and Σ is an $m \times n$ rectangular diagonal matrix with singular values on its diagonal.

We now connect this to principal component analysis (PCA). PCA is a method that uses the singular value decomposition to find a lower-dimensional representation of data that captures its most significant variance.

Example 1

Consider $A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix}$. We compute:

$$A^\top A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 4 \\ 4 & 17 \end{bmatrix}$$

The eigenvalues of $A^\top A$ are $\lambda_1 = 21$ and $\lambda_2 = 13$ giving singular values $\sigma_1 = \sqrt{21}$ and $\sigma_2 = \sqrt{13}$.
The eigenvectors of $A^\top A$ (which form the columns of V) are $v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^\top$ and $v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}^\top$.
The left singular vectors are:

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{42}} \begin{bmatrix} 5 & 5 & 0 \end{bmatrix}^\top = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{5}{\sqrt{50}} & \frac{5}{\sqrt{50}} & 0 \end{bmatrix}^T$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{\sqrt{26}} [1, 1, 4]^T$$

We need a third orthogonal left singular vector to complete U. Using the Gram-Schmidt process, we get $u_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} -\frac{1}{\sqrt{50}} & -\frac{1}{\sqrt{50}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T$.
Thus the SVD of A is:

$$A = \begin{bmatrix} \frac{5}{\sqrt{50}} & \frac{1}{\sqrt{26}} & -\frac{1}{\sqrt{50}} \\ \frac{5}{\sqrt{50}} & \frac{1}{\sqrt{26}} & -\frac{1}{\sqrt{50}} \\ 0 & \frac{4}{\sqrt{26}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{21} & 0 \\ 0 & \sqrt{13} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

If we were to apply PCA to data represented by this matrix, we would project onto the first principal component $v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, which captures the direction of maximum variance in the data.

Theorem 1 (Optimality of PCA)

The principal components obtained via SVD provide the optimal linear projection of the data in terms of minimizing reconstruction error.

Theorem 2

The action of an $m \times n$ matrix A on the unit sphere in $\mathbb{R}^n$ can be decomposed into:

1. A rotation/reflection in $\mathbb{R}^n$ (via $V^\top$).
2. A scaling along the coordinate axes by the singular values (via Σ).
3. A rotation/reflection in $\mathbb{R}^n$ (via U).

Several fundamental matrix properties are elegantly expressed via singular values.

Theorem 3

Let A be an $m \times n$ matrix with singular values $\sigma_1 \geq ... \geq \sigma_r > 0$ where $r = \text{rk}(A)$. Then:

- $\text{rk}(A) = r =$ number of non-zero singular values
- $||A||_2 = \sigma_1$ (Euclidean/ℓ_2 norm)
- $||A||_F = \sqrt{\sigma_1^2 + ... + \sigma_r^2}$ (Frobenius norm)
- $|\det(A)| = \prod_{i=1}^r \sigma_i$ (if $m = n$)
- $\ker A = \text{span}\{v_{r+1}, ..., v_n\}$
- $\text{im} A = \text{span}\{u_1, ..., u_r\}$

We shall end with a brief theorem on why PCA is optimal.

Theorem 4 (Eckart-Young-Mirsky Theorem)

The best rank-k approximation to A in the Frobenius norm is given by,

$$A_k = U_k \Sigma_k V_k^\top = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$$

where U_k, Σ_k, and V_k contain only the first K columns/singular values.