25 Truncated Singular Value Decomposition: From Geometry to Image Compression

25.1 Truncated SVD: Dimensionality Reduction with Purpose

25.1.1 Why Truncate? Preserving the Essentials

Last chapter introduces Singular Value Decomposition (SVD). In short, SVD is a foundational mathematical technique widely used in machine learning, data analysis, and signal processing.

Truncated SVD is an approximation of the original matrix that intentionally discards the smallest singular values and their associated singular vectors. Unlike full or reduced SVD, which preserve all the information and allow exact reconstruction, truncated SVD keeps only the most significant components, typically those corresponding to the largest singular values. This results in a lower-rank approximation that captures the most important patterns or structures in the data while reducing noise and dimensionality.

Although this means losing some information, the trade-off often leads to more efficient storage, faster computation, and improved performance in tasks like compression, denoising, and principal component analysis.

As shown in Figure 1, consider a data matrix $\boldsymbol{X}$ (n rows, D columns) where each row is a data sample with D features. The reduced SVD of $\boldsymbol{X}$ is expressed as

$$\boldsymbol{X}_{n\times D} = \boldsymbol{U}_{n\times D}\boldsymbol{S}_{D\times D}\boldsymbol{V}_{D\times D}^{\mathrm{T}} \tag{1}$$

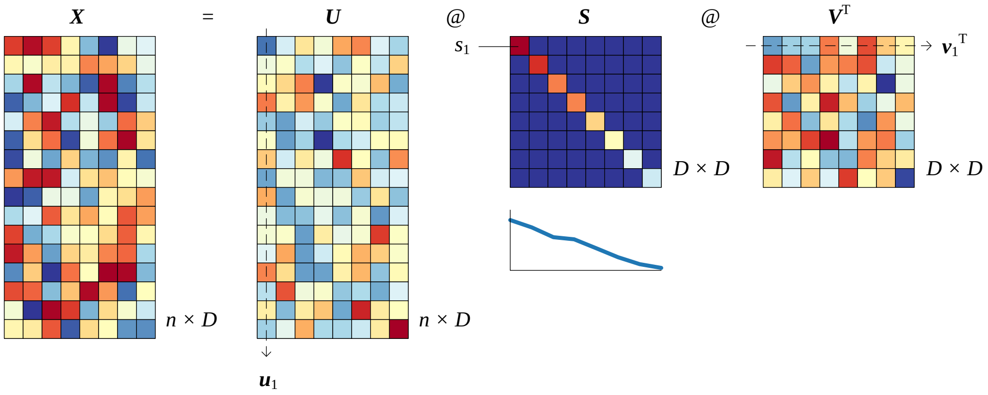

Figure 1. Reduced SVD of an $n \times D$ data matrix $\boldsymbol{X}$

25.1.2 Matrix as Sum of Rank-1 Components

Matrix multiplication can be interpreted as a sum of rank-1 matrices. Applying this to the SVD of $\boldsymbol{X}$, we obtain

$$\boldsymbol{X}_{n\times D} = \underbrace{\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D\end{bmatrix}}_{\boldsymbol{U}_{n\times D}} \underbrace{\begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix}}_{\boldsymbol{S}_{D\times D}} \underbrace{\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_D^{\mathrm{T}}\end{bmatrix}}_{\boldsymbol{V}_{D\times D}^{\mathrm{T}}} = s_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}} + s_2\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}} + \cdots + s_D\boldsymbol{u}_D\boldsymbol{v}_D^{\mathrm{T}} = \sum_{j=1}^{D} s_j\boldsymbol{u}_j\boldsymbol{v}_j^{\mathrm{T}} \tag{2}$$

Assuming $s_j > 0$, $s_j\boldsymbol{u}_j\boldsymbol{v}_j$ is a rank-1 matrix of shape $n \times D$.

Since both $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are unit vectors (i.e., have L2 norm equal to 1), the magnitude of the singular value s_j determines the importance of the corresponding principal component. Figure 2 shows the first two rank-1 matrixes.

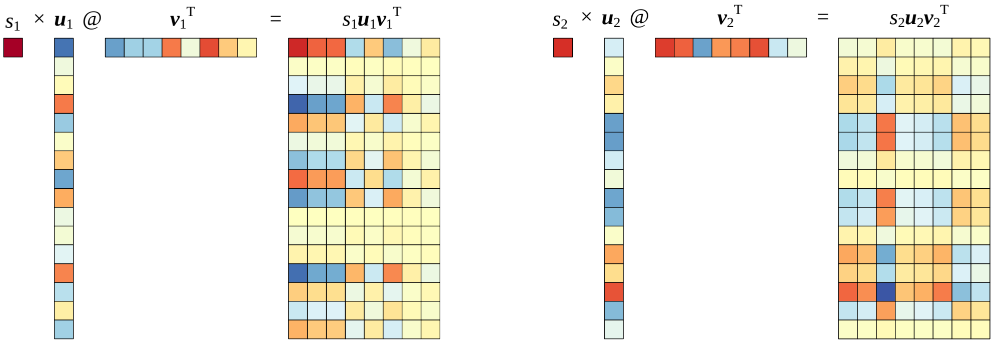
s_1 × $\boldsymbol{u}_1$ @ $\boldsymbol{v}_1^\mathrm{T}$ = $s_1\boldsymbol{u}_1\boldsymbol{v}_1^\mathrm{T}$
s_2 × $\boldsymbol{u}_2$ @ $\boldsymbol{v}_2^\mathrm{T}$ = $s_2\boldsymbol{u}_2\boldsymbol{v}_2^\mathrm{T}$

Figure 2. The first two rank-1 matrixes

25.1.3 Truncated SVD Approximation

As shown in Figure 3, we can approximate the original matrix $\boldsymbol{X}$ using only the top $\boldsymbol{p}$ singular values and corresponding vectors

$$\boldsymbol{X}_{n\times D} \approx \hat{\boldsymbol{X}}_{n\times D} = \boldsymbol{U}_{n\times p}\boldsymbol{S}_{p\times p}\left(\boldsymbol{V}_{D\times p}\right)^\mathrm{T} = \sum_{j=1}^{p} s_j\boldsymbol{u}_j\boldsymbol{v}_j^\mathrm{T} \tag{3}$$

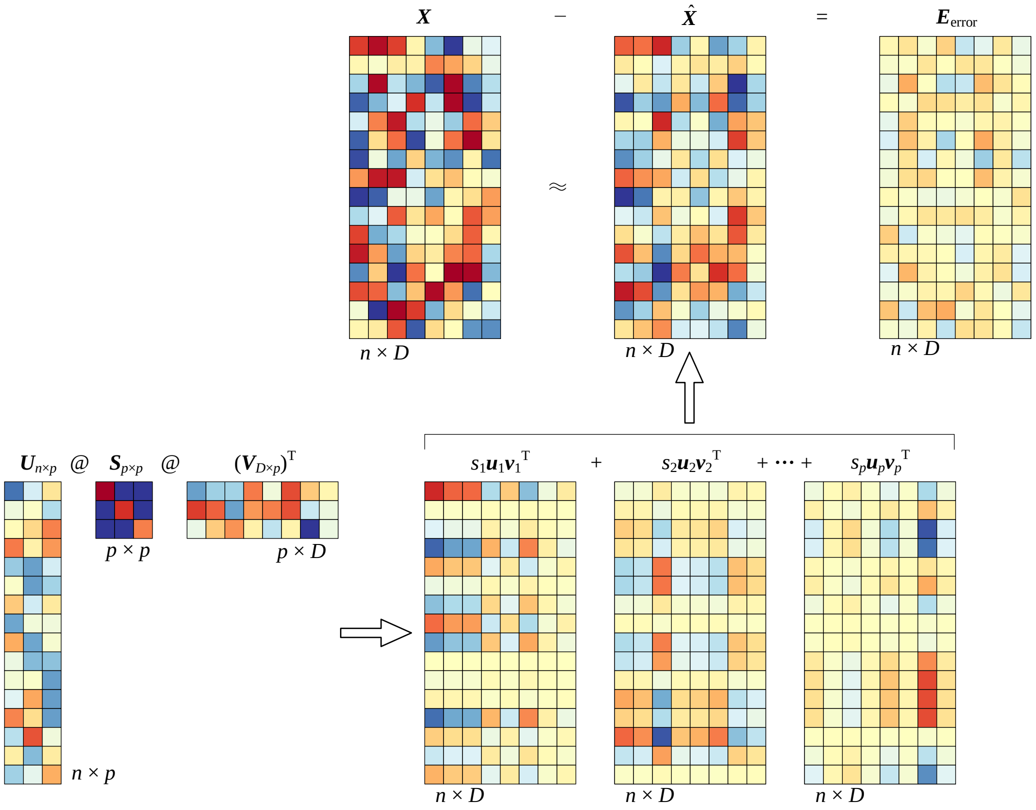
$\boldsymbol{X}$
−
$\hat{\boldsymbol{X}}$
=
$\boldsymbol{E}_{\text{error}}$
≈
$n \times D$
$n \times D$
$n \times D$
$\boldsymbol{U}_{n\times p}$ @ $\boldsymbol{S}_{p\times p}$ @ $(\boldsymbol{V}_{D\times p})^\mathrm{T}$
$p \times p$
$p \times D$
$n \times p$
$s_1\boldsymbol{u}_1\boldsymbol{v}_1^\mathrm{T}$ + $s_2\boldsymbol{u}_2\boldsymbol{v}_2^\mathrm{T}$ + ⋯ + $s_p\boldsymbol{u}_p\boldsymbol{v}_p^\mathrm{T}$
$n \times D$
$n \times D$
$n \times D$

Figure 3. Result of using only the top *p* principal components to approximate the raw data.

Also, in Figure 3, we can see the error—the difference between the original matrix $\boldsymbol{X}$ and its approximation—is the sum of the remaining components not included in the truncated SVD:

$$\boldsymbol{X}-\hat{\boldsymbol{X}}=\sum_{j=1}^{D}s_j\boldsymbol{u}_j\boldsymbol{v}_j^{\mathrm{T}}-\sum_{j=1}^{p}s_j\boldsymbol{u}_j\boldsymbol{v}_j^{\mathrm{T}}=\sum_{j=p+1}^{D}s_j\boldsymbol{u}_j\boldsymbol{v}_j^{\mathrm{T}} \tag{4}$$

25.2 Optimization Behind SVD

25.2.1 Finding Principal Directions

With the geometric and algebraic foundations laid above, we now explore the optimization problem underlying SVD.

As illustrated in Figure 4, we aim to find a unit vector $\boldsymbol{v}$ such that the projection of the data matrix $\boldsymbol{X}$ onto $\boldsymbol{v}$, i.e., $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{v}$, has the maximum L2 norm:

$$\begin{aligned}&\underset{\boldsymbol{v}}{\arg\max}\,\|\boldsymbol{X}\boldsymbol{v}\|\\&\text{subject to: }\|\boldsymbol{v}\|=1\end{aligned} \tag{5}$$

This problem searches for the direction in which the projected data $\boldsymbol{y}$ exhibits the greatest variance. The maximum value of this expression is precisely the largest singular value s_1, and the corresponding $\boldsymbol{v}_1$ is the first right singular vector of $\boldsymbol{X}$.

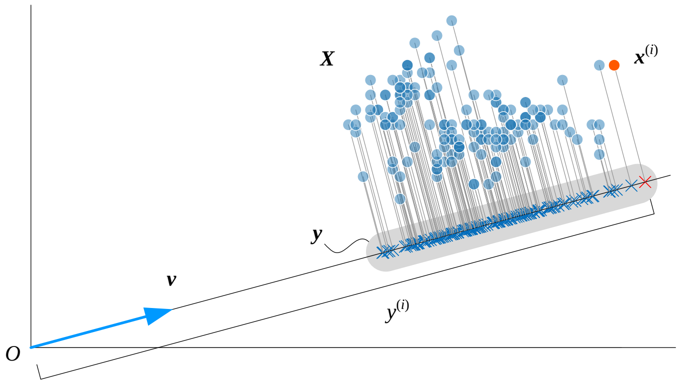

Figure 4. The raw data projected to a unit vector $\boldsymbol{v}$

25.2.2 Rayleigh Quotient Perspective

The optimization problem described above can also be expressed through the **Rayleigh quotient**

$$R(\boldsymbol{v})=\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \tag{6}$$

For any non-zero $\boldsymbol{v}$, this quotient achieves its maximum when $\boldsymbol{v} = \boldsymbol{v}_1$, the dominant eigenvector of the Gram matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

Rayleigh quotient is a powerful concept that connects matrix transformations, quadratic forms, and eigenvalues.

The standard Rayleigh quotient for a symmetric matrix $\boldsymbol{A}$ is defined as:

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} \tag{7}$$

As shown in Figure 5, both the numerator and the denominator of the Rayleigh quotient are **quadratic forms** in the vector $\boldsymbol{x}$.

The numerator $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$ represents how the matrix $\boldsymbol{A}$ transforms and scales the vector $\boldsymbol{x}$, while the denominator $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ is simply the squared **L2 norm** of $\boldsymbol{x}$.

Since the denominator cannot be zero, the Rayleigh quotient is undefined for the zero vector — all components of **x** cannot simultaneously be zero.

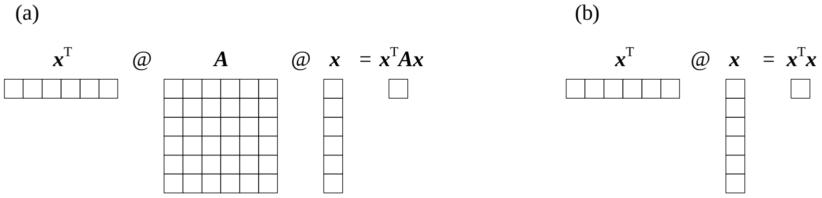
(a)
$\boldsymbol{x}^{\mathrm{T}}$
@
$\boldsymbol{A}$
@
$\boldsymbol{x}$
$= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$
(b)
$\boldsymbol{x}^{\mathrm{T}}$
@
$\boldsymbol{x}$
$= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$

Figure 5. Rayleigh Quotient: Numerator and Denominator as Matrix Quadratic Forms

The values of the Rayleigh quotient are bounded between the smallest and largest eigenvalues of $\boldsymbol{A}$:

$$\lambda_{\min} \le R(\boldsymbol{x}) \le \lambda_{\max} \tag{8}$$

This means that as x changes direction, $R(\boldsymbol{x})$ varies smoothly between these two extremes. In essence, the Rayleigh quotient is a **multivariable function** that depends only on the **direction** of $\boldsymbol{x}$, not its magnitude.

To build intuition, consider a two-dimensional Rayleigh quotient of the form with

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{9}$$

Thus the Rayleigh quotient is bivariate function

$$R(x_1, x_2) = \frac{2x_1^2 + 2x_2^2 + 2x_1x_2}{x_1^2 + x_2^2}. \tag{10}$$

The denominator ensures that the function is undefined at the origin (0,0), so its **domain** excludes that point. The surface and contour plots of this function are shown in Figure 6, where we can already identify regions corresponding to its maximum and minimum values.

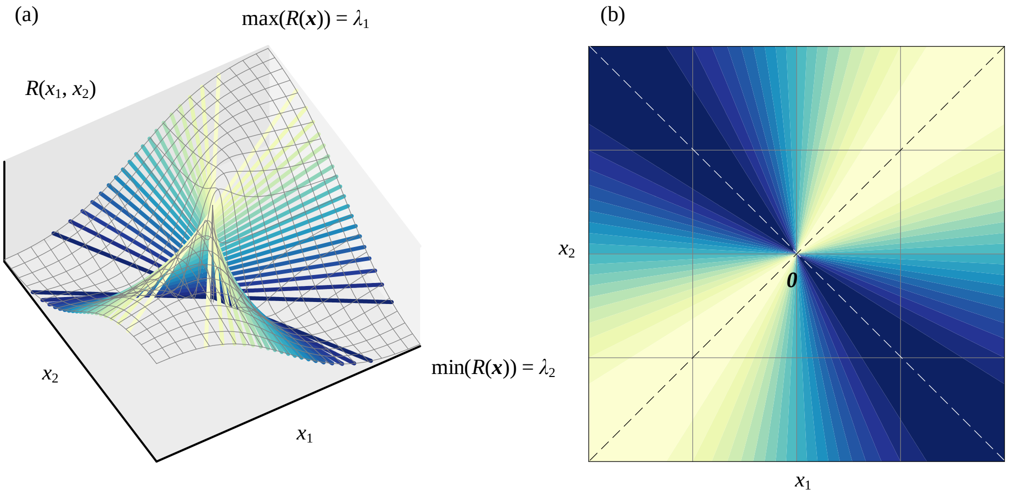

Figure 6. Surface and Contour Plots of a Two-Dimensional Rayleigh Quotient

Interestingly, the contour plot of $R(x_1, x_2)$ reveals a radial symmetry—its values depend only on the direction of $\boldsymbol{x}$, not its length. This naturally leads us to use **polar coordinates**:

$$\begin{cases} x_1 = r\cos\theta \\ x_2 = r\sin\theta \end{cases} \tag{11}$$

Substituting these expressions into $R(x_1, x_2)$, we find that the radius r cancels out completely, leaving a function that depends only on the angle θ

$$\begin{aligned} R(\theta) &= \frac{2r^2\cos^2\theta + 2r^2\sin^2\theta + 2r^2\cos\theta\sin\theta}{r^2} \\ &= 2\cos^2\theta + 2\sin^2\theta + 2\cos\theta\sin\theta \\ &= 2 + \sin(2\theta) \end{aligned} \tag{12}$$

This confirms our observation: the Rayleigh quotient in 2D depends purely on **direction**. The resulting trigonometric curve, shown in Figure 7, clearly indicates the angles where the quotient attains its maximum and minimum.

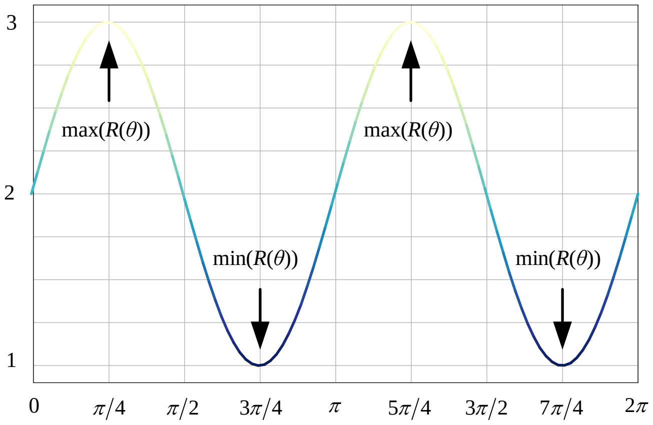

Figure 7. Trigonometric Representation of the 2D Rayleigh Quotient

Because the denominator $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ can be written as $\|\boldsymbol{x}\|_2^2$ we can express the quotient as:

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2} = \left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)^{\mathrm{T}} \boldsymbol{A}\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right) = \hat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{A}\hat{\boldsymbol{x}} \tag{13}$$

Let $\hat{\boldsymbol{x}}$ be the **unit vector** (the direction of $\boldsymbol{x}$). The endpoints of all such unit vectors lie on the **unit circle**, so we can visualize the Rayleigh quotient directly on that circle, as shown in Figure 8.

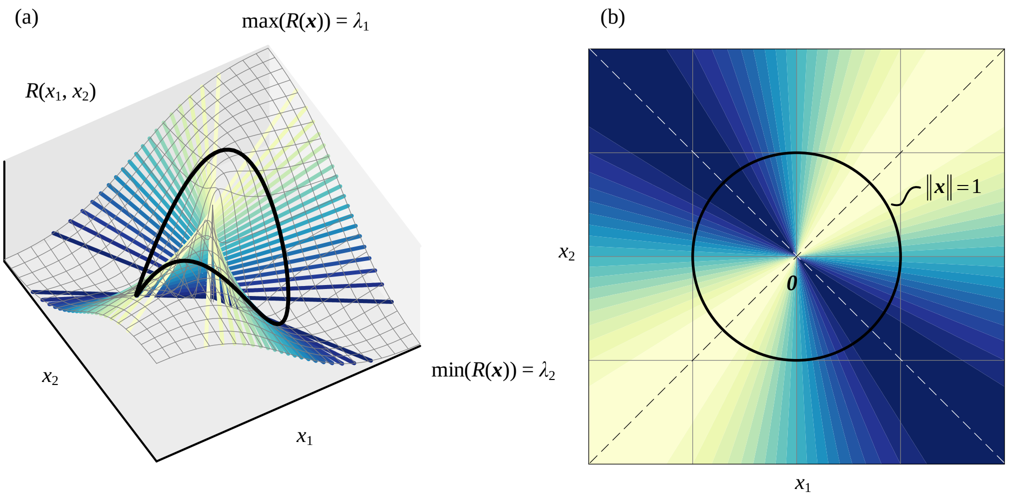

Figure 8. Contour Plot of the Rayleigh Quotient Observed on the Unit Circle

Since $\boldsymbol{A}$ is a symmetric real matrix, we perform the **spectral decomposition** of matrix $\boldsymbol{A}$,

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = \underbrace{\begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}}_{\boldsymbol{V}} \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}}_{\boldsymbol{V}^{\mathrm{T}}} \tag{14}$$

and write $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2]$, then $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ correspond to the directions where the Rayleigh quotient achieves its **maximum** (λ_1) and **minimum** (λ_2) values, respectively (Figure 9).

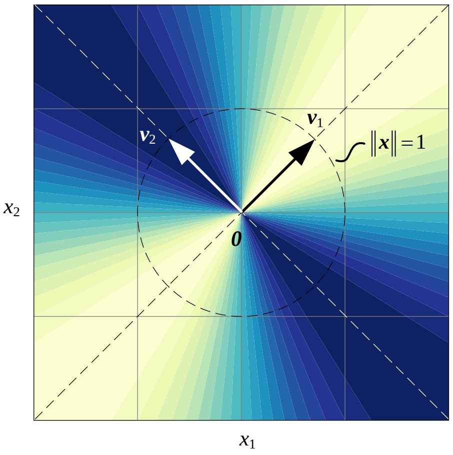

Figure 9. Directions of Maximum and Minimum Rayleigh Quotient

Along $\boldsymbol{v}_1$, we have:

$$\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_1=\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{v}_1=\boldsymbol{v}_1^{\mathrm{T}}\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2\end{bmatrix}\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\end{bmatrix}\boldsymbol{v}_1=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2\end{bmatrix}\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1\\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_1\end{bmatrix}=\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix}=\lambda_1 \quad (15)$$

and along $\boldsymbol{v}_2$,

$$\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_2=\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{v}_2=\boldsymbol{v}_2^{\mathrm{T}}\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2\end{bmatrix}\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\end{bmatrix}\boldsymbol{v}_2=\begin{bmatrix}\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_1 & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_2\end{bmatrix}\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2\\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_2\end{bmatrix}=\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix}=\lambda_2 \quad (16)$$

Thus, the Rayleigh quotient achieves its extreme values along the eigenvector directions of $\boldsymbol{A}$.

Just as the endpoints of all 2D unit vectors lie on a unit circle, the endpoints of all 3D unit vectors lie on a unit sphere. Hence, we can visualize a three-dimensional Rayleigh quotient over the surface of a sphere.

Consider the symmetric 3×3 matrix:

$$\boldsymbol{A}=\begin{bmatrix}1 & 0.5 & 1\\ 0.5 & 2 & -0.2\\ 1 & -0.2 & 1\end{bmatrix} \quad (17)$$

Its corresponding Rayleigh quotient is a three-variable function,

$$f\left(x_1,x_2,x_3\right)=\frac{x_1^2+2x_2^2+x_3^2+x_1x_2+2x_1x_3-0.4x_2x_3}{x_1^2+x_2^2+x_3^2} \quad (18)$$

As illustrated in Figure 10, we can view this function on the unit sphere. The matrix **A** has eigenvalues approximately 2.272, 1.844, and −0.117, meaning that the Rayleigh quotient attains its maximum and minimum values at 2.272, and −0.117.

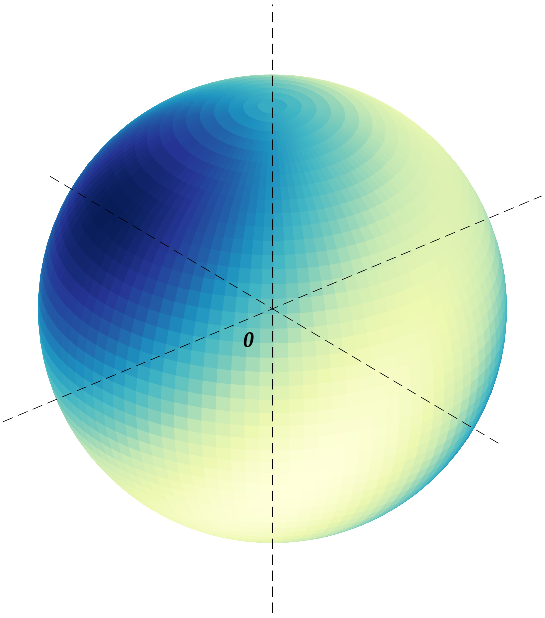

Figure 10. Visualizing the 3D Rayleigh Quotient on the Unit Sphere

To better capture the full surface of the 3D Rayleigh quotient, imagine wrapping the unit sphere into a globe. Using spherical coordinates, the resulting "Rayleigh Earth" map in Figure 11 displays the quotient's values in a geographic style—highlighting the angular dependence of $R(\boldsymbol{x})$ in terms of latitude and longitude.

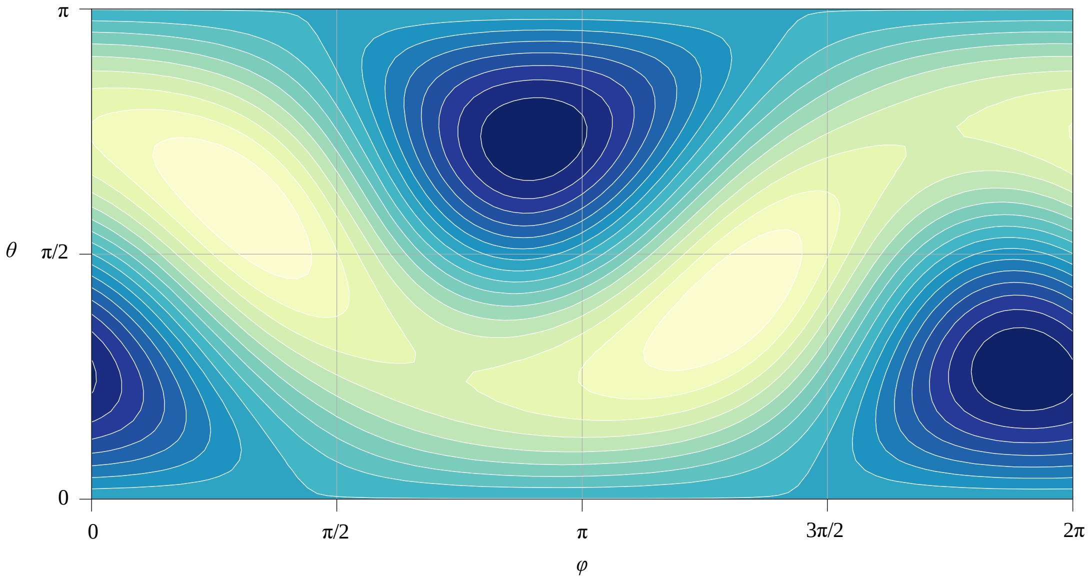

Figure 11. "Rayleigh Earth": Viewing the 3D Rayleigh Quotient in Latitude–Longitude Coordinates

25.3 Image Compression and Reconstruction

25.3.1 Preparing the Image

In this section, we illustrate the power of truncated SVD using an image compression and reconstruction task. The subject is an iris photograph taken by the author. After converting the image to grayscale, as shown in Figure 12, each pixel is represented by a value in the range [0, 1], making the image essentially a matrix of real numbers.

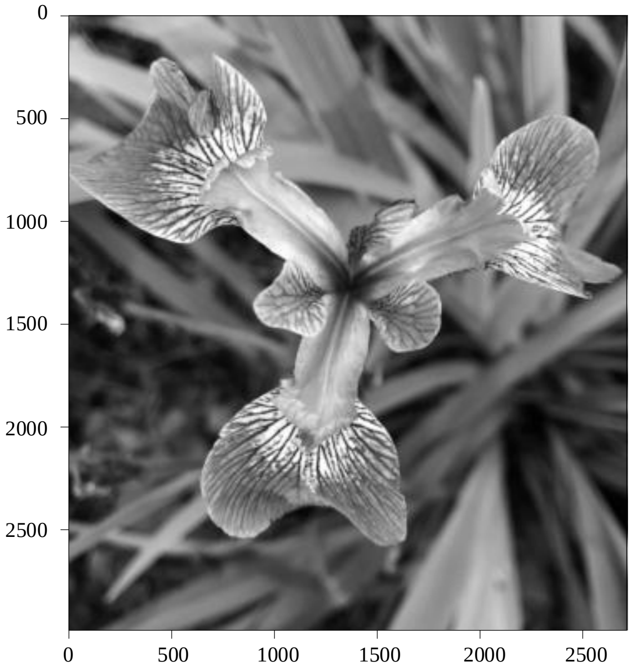

Figure 12. Iris photograph, grayscale. Figure generated by Ch25_01_Truncated_SVD.ipynb.

25.3.2 Visualizing Singular Values

Figure 13 show how singular values decay with the number of principal components.

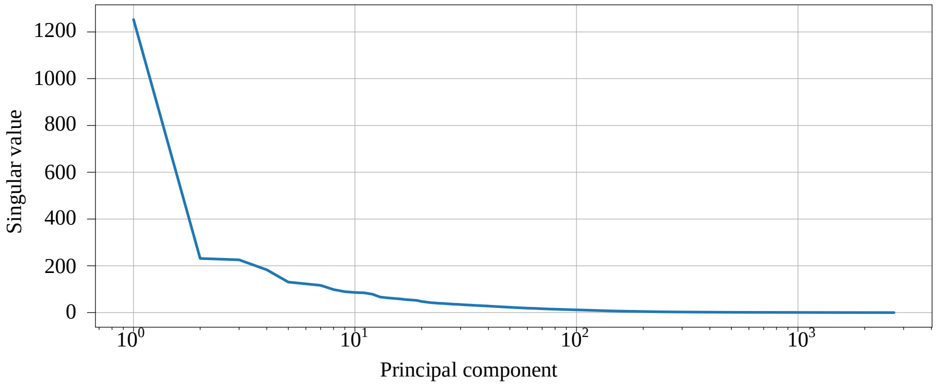

Figure 13. The singular value varies with the principal component. Figure generated by Ch25_01_Truncated_SVD.ipynb.

25.3.3 Reconstruction with Truncated SVD

Figure 14 shows the image reconstructed using only the two singular values (rank-2 approximation). The reconstruction is extremely coarse—while the general structure is preserved, fine details of the iris are not discernible.

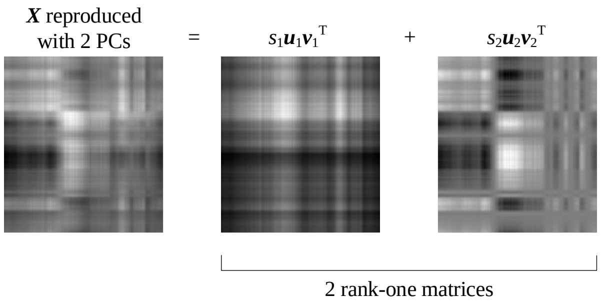
$\boldsymbol{X}$ reproduced with 2 PCs
=
$s_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}$
+
$s_2\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}$
2 rank-one matrices

Figure 14. The first two ranks are superimposed on the first matrix. Figure generated by Ch25_01_Truncated_SVD.ipynb.

Figure 15 show the reconstruction using the first four principal components (rank-4 approximation). The image becomes more recognizable, and major contours start to emerge.

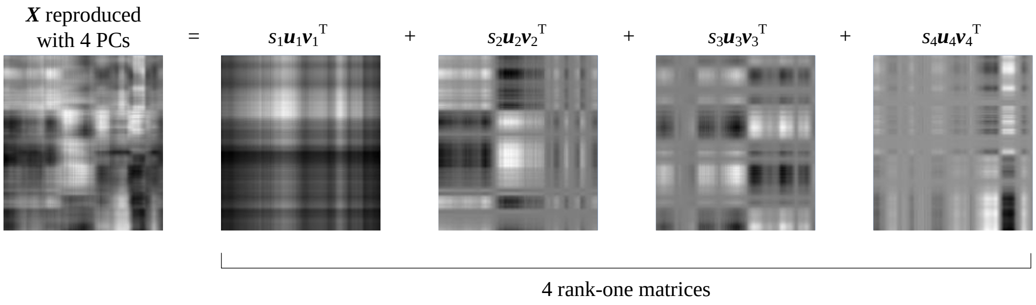
$\boldsymbol{X}$ reproduced with 4 PCs
=
$s_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}$
+
$s_2\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}$
+
$s_3\boldsymbol{u}_3\boldsymbol{v}_3^{\mathrm{T}}$
+
$s_4\boldsymbol{u}_4\boldsymbol{v}_4^{\mathrm{T}}$
4 rank-one matrices

Figure 15. The first 4 ranks are superimposed on the first matrix. Figure generated by Ch25_01_Truncated_SVD.ipynb.

In Figure 16, 15 singular values are used (rank-15 approximation). The iris becomes clearly visible, and the structure of the image is largely recovered. This implies the original image can be well-approximated with just 15 rank-1 matrices.

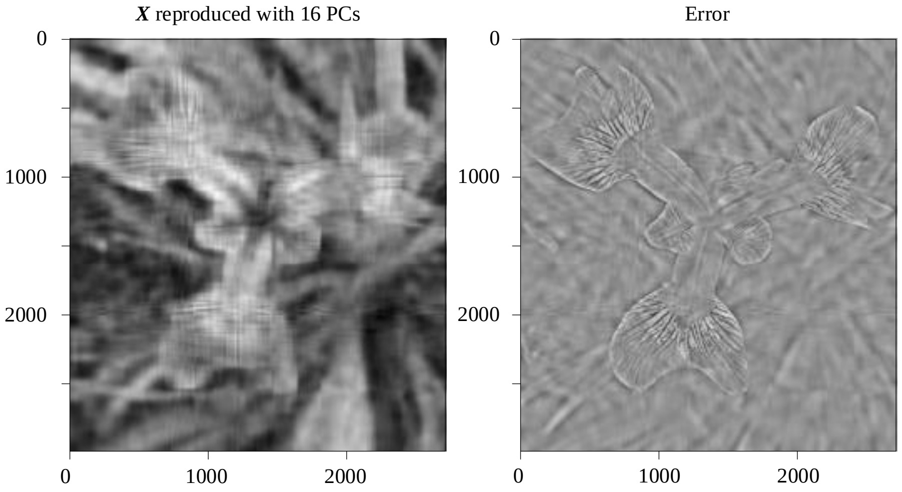

Figure 16. Use the first 16 main elements to restore the iris photo. Figure generated by Ch25_01_Truncated_SVD.ipynb.

This method of dimensionality reduction via truncated SVD is widely used in computer vision, especially in face recognition. One prominent application is the eigenfaces approach, where principal components of face images are used to form a low-dimensional feature space. These "feature faces" are linear combinations of original face images in the training set and capture the most important variations across individuals.

25.4 Conclusion

This chapter introduces Singular Value Decomposition (SVD), a powerful mathematical tool used for analyzing and simplifying data. It begins by explaining the full and reduced forms of SVD, which decompose a matrix into orthogonal components that reveal its internal structure.

From a geometric perspective, SVD can be viewed as a sequence of rotations and scaling operations that transform data across dimensions.

The chapter then focuses on truncated SVD, a form of dimensionality reduction that keeps only the most important components, making it useful for compressing data and removing noise. Through visual examples, including the reconstruction of a grayscale iris photograph, the chapter demonstrates how truncated SVD can significantly reduce the size of an image while preserving its essential features.