4

Matrix Decompositions

In Chapters 2 and 3, we studied ways to manipulate and measure vectors, projections of vectors and linear mappings. Mappings and transformations of vectors can be conveniently described as operations performed on matrices. Moreover, data is often represented in matrix form as well, for example where the rows of the matrix represent different instances of the data (for example people) and the columns describe different features of the data (for example weight, height and socio-economic status). In this chapter we present three aspects of matrices: how to summarize matrices, how matrices can be decomposed, and how these decompositions can be used to consider matrix approximations.

We first consider methods that allow us to describe matrices with just a few numbers that characterize the overall properties of matrices. We will do this in the sections on determinants (Section 4.1) and eigenvalues (Section 4.2) for the important special case of square matrices. These characteristic numbers have important mathematical consequences and allow us to quickly grasp what useful properties a matrix has. From here we will proceed to matrix decomposition methods: An analogy for matrix decomposition is the factoring of numbers, such as the factoring of 21 into prime numbers 7×3. For this reason matrix decomposition is also often referred to as *matrix factorization*. Matrix decompositions are used to interpret a matrix using a different representation using factors of interpretable matrices.

matrix factorization

We will first cover a square-root-like operation for matrices called Cholesky decomposition (Section 4.3) for symmetric, positive definite matrices. From here we will look at two related methods for factorizing matrices into canonical forms. The first one is known as matrix diagonalization (Section 4.4), which allows us to represent the linear mapping using a diagonal transformation matrix if we choose an appropriate basis. The second method, singular value decomposition (Section 4.5), extends this factorization to non-square matrices, and it is considered one of the fundamental concepts in linear algebra. These decompositions are helpful as matrices representing numerical data are often very large and hard to analyze. We conclude the chapter with a systematic overview of the types of matrices and the characteristic properties that distinguish them in form of a matrix taxonomy (Section 4.7).

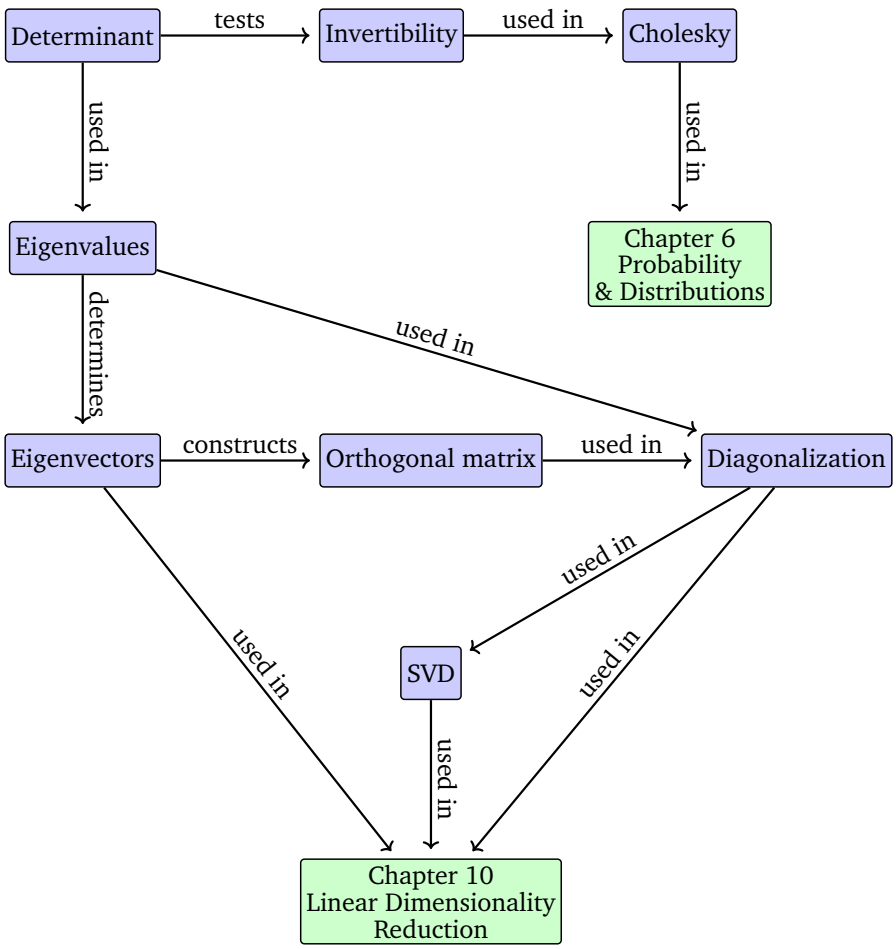


Figure 4.1 A mind map of the concepts introduced in this chapter, along with when they are used in other parts of the book.

The methods that we cover in this chapter will become important in both subsequent mathematical chapters, such as Chapter 6 but also in applied chapters, such as dimensionality reduction in Chapters 10 or density estimation in Chapter 11. This chapter's overall structure is depicted in the mind map of Figure 4.1.

4.1 Determinant and Trace

Determinants are important concepts in linear algebra. A determinant is a mathematical object in the analysis and solution of systems of linear equations. Determinants are only defined for square matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, that is matrices with the same number of rows and columns. In this book we write this as $\det(\boldsymbol{A})$ (some textbooks may use $|\boldsymbol{A}|$, which we find confusing in terms of notation with the absolute value). However, we will use the straight lines when we write out the full matrix. Recall that a_{ij} is

the element in the i^{th} row and j^{th} column of a matrix $\boldsymbol{A}$. Then we write

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} . \tag{4.1}$$

The *determinant* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a function that maps $\boldsymbol{A}$ onto a real number. Before providing a definition of the determinant for general $n \times n$ matrices let us look at some motivating examples, and define determinants for some special matrices.

determinant

Example 4.1 (Testing for Matrix Invertibility)
Let us begin with exploring if a square matrix $\boldsymbol{A}$ is invertible (see Section 2.2.2). For the smallest cases, we already know when a matrix is invertible. If $\boldsymbol{A}$ is a 1×1 matrix, i.e., it is a scalar number, then $\boldsymbol{A} = a \implies \boldsymbol{A}^{-1} = \frac{1}{a}$. Thus $a\,\frac{1}{a} = 1$ holds, if and only if $a \neq 0$.

For the case of 2×2 matrices, by the definition of the inverse (Definition 2.3), we know that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ and thus can write that the inverse of $\boldsymbol{A}^{-1}$ is (from Equation 2.24)

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} . \tag{4.2}$$

Therefore, $\boldsymbol{A}$ is invertible if and only if

$$a_{11}a_{22} - a_{12}a_{21} \neq 0 . \tag{4.3}$$

This quantity is the determinant of $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, that is

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} . \tag{4.4}$$

The example above points already at the relationship between determinants and the existence of inverse matrices. The next theorem states the same result for $n \times n$ matrices.

Theorem 4.1 *For any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ it holds that $\boldsymbol{A}$ is invertible if and only if $\det(\boldsymbol{A}) \neq 0$.*

We have explicit (closed form) expressions for determinants of small matrices in terms of the elements of the matrix. For $n = 1$,

$$\det(\boldsymbol{A}) = \det(a_{11}) = a_{11} . \tag{4.5}$$

For $n = 2$,

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} , \tag{4.6}$$

which we have observed in the example above. For $n = 3$ (known as Sarrus' rule),

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} \,. \tag{4.7}$$

For a memory aid of the product terms in Sarrus' rule, try tracing the elements of the triple products in the matrix. We call a square matrix $\boldsymbol{A}$ a *upper triangular matrix* if $a_{ij} = 0$ for $i > j$, that is the matrix is zero below its diagonal. Analogously, we define a *lower triangular matrix* as a matrix with zeros above its diagonal. For an upper/lower triangular matrix $\boldsymbol{A}$, the determinant is the product of the diagonal elements:

upper triangular matrix
lower triangular matrix

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} a_{ii}. \tag{4.8}$$

The determinant is the signed volume of the parallelepiped formed by the columns of the matrix.

Example 4.2 (Determinants as Measures of Volume)
The notion of a determinant is natural when we consider it as a mapping from a set of n vectors spanning an object in $\mathbb{R}^n$. It turns out that the determinant is then the signed volume of an n-dimensional parallelepiped formed by columns of a matrix $\boldsymbol{A}$.

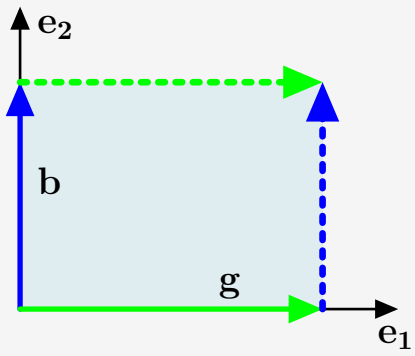


Figure 4.2 Determinants can measure areas spanned by vectors. The area A of the parallelogram (shaded region) spanned by the vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ is given by the determinant $|\det([\boldsymbol{b}, \boldsymbol{g}])|$.

For $n = 2$ the columns of the matrix form a parallelogram. As the angle between vectors gets smaller the area of a parallelogram shrinks, too. Figure 4.2 illustrates this setting. Assume two linearly independent vectors $\boldsymbol{b}, \boldsymbol{g}$ that form the columns of a matrix $\boldsymbol{A} = [\boldsymbol{b}, \boldsymbol{g}]$. Then, the absolute value of the determinant of $\boldsymbol{A}$ is the area of the parallelogram with vertices $\mathbf{0}, \boldsymbol{b}, \boldsymbol{g}, \boldsymbol{b} + \boldsymbol{g}$. In particular, if the two vectors $\boldsymbol{b}, \boldsymbol{g}$ were linearly dependent so that $\boldsymbol{b} = \lambda \boldsymbol{g}$ for some $\lambda \in \mathbb{R}$ they no longer form a two-dimensional parallelogram. Therefore, the corresponding area is 0. On the contrary, if $\boldsymbol{b}, \boldsymbol{g}$ were linearly independent and lie along the canonical

coordinate axes $\boldsymbol{e}_1, \boldsymbol{e}_2$ then they would reduce to $\boldsymbol{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$ and $\boldsymbol{g} = \begin{bmatrix} 0 \\ g \end{bmatrix}$ and the determinant

$$\begin{vmatrix} b & 0 \\ 0 & g \end{vmatrix} = bg - 0 = bg \tag{4.9}$$

becomes the familiar formula: area = height × length.

The sign of the determinant measures the orientation of the spanning vectors $\boldsymbol{b}, \boldsymbol{g}$ with respect to the standard coordinate system $\boldsymbol{e}_1, \boldsymbol{e}_2$. In our figure, flipping the spanning order to $\boldsymbol{g}, \boldsymbol{b}$ swaps the columns of $\boldsymbol{A}$ and reverses the orientation of the shaded surface A.

This intuition extends to higher dimensions. In $\mathbb{R}^3$, we consider three vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g} \in \mathbb{R}^3$ spanning the edges of a parallelepiped, i.e., a solid with faces that are parallel parallelograms (see Figure 4.3). The absolute value of the determinant of the 3×3 matrix $[\boldsymbol{r},\ \boldsymbol{b},\ \boldsymbol{g}]$ is the volume of the solid. Thus, the determinant acts as a function that measures the signed volume formed by column vectors composed in a matrix.

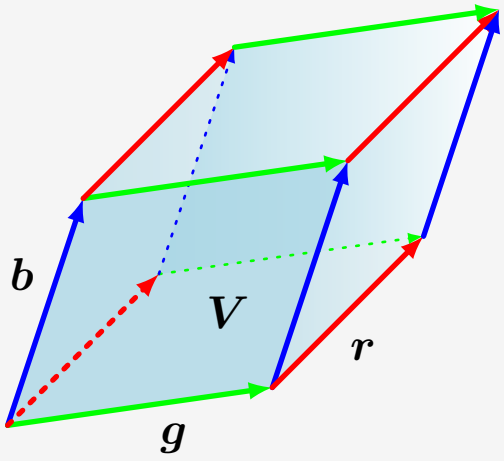


Figure 4.3 Determinants can measure volumes spanned by vectors. The volume of the parallelepiped (shaded volume) spanned by vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}$ is given by the determinant $|\det([\boldsymbol{r},\ \boldsymbol{b},\ \boldsymbol{g}])|$.

Consider the three linearly independent vectors $\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b} \in \mathbb{R}^3$ given as

$$\boldsymbol{r} = \begin{bmatrix} 2 \\ 0 \\ -8 \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} 6 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}. \tag{4.10}$$

$$\boldsymbol{A} = [\boldsymbol{r},\ \boldsymbol{g},\ \boldsymbol{b}] = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 1 & 4 \\ -8 & 0 & -1 \end{bmatrix}. \tag{4.11}$$

Therefore, the volume is given as

$$V = |\det(\boldsymbol{A})| = 186\,. \tag{4.12}$$

Computing the determinant of an $n \times n$ matrix requires a general algorithm to solve the cases for $n > 3$, which we are going to explore in the following. The theorem below reduces the problem of computing the determinant of an $n \times n$ matrix to computing the determinant of $(n-1) \times (n-1)$ matrices. By recursively applying the following Laplace expansion we can therefore compute determinants of $n \times n$ matrices by ultimately computing determinants of 2×2 matrices.

Theorem 4.2 (Laplace Expansion) *Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Then, for all $j = 1, \ldots, n$:*

$\det(\boldsymbol{A}_{k,j})$ is called a *minor* and $(-1)^{k+j} \det(\boldsymbol{A}_{k,j})$ a *cofactor*.

1 Expansion along column j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} \det(\boldsymbol{A}_{k,j}) \,. \tag{4.13}$$

2 Expansion along row j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{jk} \det(\boldsymbol{A}_{j,k}) \,. \tag{4.14}$$

Here $\boldsymbol{A}_{k,j} \in \mathbb{R}^{(n-1) \times (n-1)}$ is the submatrix of $\boldsymbol{A}$ that we obtain when deleting row k and column j.

Example 4.3 (Laplace Expansion)
Let us compute the determinant of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.15}$$

using the Laplace expansion along the first row. By applying (4.14) we obtain

$$\begin{aligned} \begin{vmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix} &= (-1)^{1+1} \cdot 1 \begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} \\ &\quad + (-1)^{1+2} \cdot 2 \begin{vmatrix} 3 & 2 \\ 0 & 1 \end{vmatrix} + (-1)^{1+3} \cdot 3 \begin{vmatrix} 3 & 1 \\ 0 & 0 \end{vmatrix} \,. \end{aligned} \tag{4.16}$$

Then we can use (4.6) to compute the determinants of all 2×2 matrices and obtain.

$$\det(\boldsymbol{A}) = 1(1-0) - 2(3-0) + 3(0-0) = -5 \,.$$

For completeness we can compare this result to computing the determinant using Sarrus' rule (4.7):

$$\det(\boldsymbol{A}) = 1 \cdot 1 \cdot 1 + 3 \cdot 0 \cdot 3 + 0 \cdot 2 \cdot 2 - 0 \cdot 1 \cdot 3 - 1 \cdot 0 \cdot 2 - 3 \cdot 2 \cdot 1 = 1 - 6 = -5 \,. \tag{4.17}$$

For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the determinant exhibits the following properties:

- The determinant of a product is the product of the determinants, $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- Determinants are invariant to transposition $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^\top)$.
- If $\boldsymbol{A}$ is regular (Section 2.2.2) then $\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$
- Similar matrices (Defintion 2.21) possess the same determinant. Therefore, for a linear mapping $\Phi : V \to V$ all transformation matrices $\boldsymbol{A}_\Phi$ of Φ have the same determinant. Thus, the determinant is invariant to the choice of basis of a linear mapping.
- Adding a multiple of a column/row to another one does not change $\det(\boldsymbol{A})$.
- Multiplication of a column/row with $\lambda \in \mathbb{R}$ scales $\det(\boldsymbol{A})$ by λ. In particular, $\det(\lambda\boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$.
- Swapping two rows/columns changes the sign of $\det(\boldsymbol{A})$.

Because of the last three properties, we can use Gaussian elimination (see Section 2.1) to compute $\det(\boldsymbol{A})$ by bringing $\boldsymbol{A}$ into row-echelon form. We can stop Gaussian elimination when we have $\boldsymbol{A}$ in a triangular form where the elements below the diagonal are all 0. Recall from Equation (4.8) that the determinant is then the product of the diagonal elements.

Theorem 4.3 *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has $\det(\boldsymbol{A}) \neq 0$ if and only if $\mathrm{rk}\boldsymbol{A} = n$. In other words a square matrix is invertible if and only if it is full rank.*

When mathematics was mainly performed by hand, the determinant calculation was considered an essential way to analyze matrix invertibility. However, contemporary approaches in machine learning use direct numerical methods that superseded the explicit calculation of the determinant. For example, in Chapter 2 we learned that inverse matrices can be computed by Gaussian elimination. Gaussian elimination can thus be used to compute the determinant of a matrix.

Determinants will play an important theoretical role for the following sections, especially when we learn about eigenvalues and eigenvectors (Section 4.2) through the characteristic polynomial of a matrix.

Definition 4.4 The *trace* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a linear function denoted by $\mathrm{tr}(\boldsymbol{A})$ and defined as trace

$$\mathrm{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} a_{ii}\,, \tag{4.18}$$

in other words, the trace is the sum of the diagonal elements of $\boldsymbol{A}$.

Remark For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ the trace satisfies the following properties:

1 $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
2 $\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\mathrm{tr}(\boldsymbol{A})\,, \quad \alpha \in \mathbb{R}$

3 $\operatorname{tr}(\boldsymbol{I}_n) = n$
4 $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$

It can be shown that only one function satisfies these four properties together – the trace (Gohberg et al., 2012). ♢

Remark The properties of the trace of matrix products are more general:

cyclic permutations

- The trace is invariant under *cyclic permutations*, i.e.,

$$\operatorname{tr}(\boldsymbol{AKL}) = \operatorname{tr}(\boldsymbol{KLA}) \tag{4.19}$$

  for matrices $\boldsymbol{A} \in \mathbb{R}^{a\times k}, \boldsymbol{K} \in \mathbb{R}^{l\times l}, \boldsymbol{L} \in \mathbb{R}^{l\times a}$. This property generalizes to products of arbitrarily many matrices.
- As a special case of (4.19) it follows that the trace is invariant under permutations of two non-square matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$:

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})\,. \tag{4.20}$$

  In particular, this means that for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\operatorname{tr}(\boldsymbol{xy}^\top) = \operatorname{tr}(\boldsymbol{y}^\top\boldsymbol{x}) = \boldsymbol{y}^\top\boldsymbol{x} \in \mathbb{R}\,. \tag{4.21}$$

♢

Remark Given some linear map $\Phi : V \to V$, we define the trace of this map by considering the trace of matrix representation of ϕ. We need to choose a basis for V and describe Φ as a matrix $\boldsymbol{A}$ relative to this basis, and taking the trace of this square matrix. Assume that $\boldsymbol{B}$ is a transformation matrix between bases of V. Then, we can write

$$\operatorname{tr}(\boldsymbol{BAB}^{-1}) = \operatorname{tr}(\boldsymbol{B}^{-1}\boldsymbol{BA}) = \operatorname{tr}(\boldsymbol{IA}) = \operatorname{tr}(\boldsymbol{A})\,. \tag{4.22}$$

Thus, while matrix representations of linear mappings are basis dependent its trace is independent of the basis. ♢

The trace is useful in certain classes of machine learning models where data is fitted using linear regression. The trace captures model complexity and can be used to compare between models (a more principled foundation for model comparison is discussed in detail in Section 8.5).

In this section, we covered determinants and traces as functions characterizing a square matrix. Taking together our understanding of determinants and traces we can now define an important equation describing a matrix $\boldsymbol{A}$ in terms of a polynomial, which we will use extensively in the following sections.

Definition 4.5 (Characteristic Polynomial) For $\lambda \in \mathbb{R}$ and a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$

$$p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \tag{4.23}$$
$$= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n\,, \tag{4.24}$$

characteristic polynomial

$c_0, \ldots, c_{n-1} \in \mathbb{R}$, is the *characteristic polynomial* of $\boldsymbol{A}$. In particular,

$$c_0 = \det(\boldsymbol{A})\,, \tag{4.25}$$
$$c_{n-1} = (-1)^{n-1}\mathrm{tr}(\boldsymbol{A})\,. \tag{4.26}$$

The characteristic polynomial will allow us to compute eigenvalues and eigenvectors, covered in the next section.

4.2 Eigenvalues and Eigenvectors

We will now get to know a new way to characterize a matrix and its associated linear mapping. Let us recall from Section 2.7.1 that every linear mapping has a unique transformation matrix given an ordered basis. We can interpret linear mappings and their associated transformation matrices by performing an "Eigen" analysis. *Eigen* is a German word meaning "characteristic", "self" or "own". As we will see the eigenvalues of a linear mapping will tell us how a special set of vectors, the eigenvectors, are transformed by the linear mapping.

Definition 4.6 Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\boldsymbol{A}$ and a nonzero $\boldsymbol{x} \in \mathbb{R}^n$ is the corresponding *eigenvector* of $\boldsymbol{A}$ if

eigenvalue
eigenvector

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}\,. \tag{4.27}$$

We call this the *eigenvalue equation*.

eigenvalue equation

Remark In linear algebra literature and software, it is a often a convention that eigenvalues are sorted in descending order, so that the largest eigenvalue and associated eigenvector are called the first eigenvalue and its associated eigenvector, and the second largest called the second eigenvalue and its associated eigenvector, and so on. However textbooks and publications may have different or no notion of orderings. We do not want to presume an ordering in our book. ♢

Definition 4.7 (Collinearity & Codirection) Two vectors that point in the same direction are called *codirected*. Two vectors are *collinear* if they point in the same or the opposite direction.

codirected
collinear

Remark (Non-uniqueness of Eigenvectors) If $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ associated with eigenvalue λ then for any $c \in \mathbb{R}\setminus\{0\}$ it holds that $c\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ with the same eigenvalue since

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}\boldsymbol{x} = c\lambda\boldsymbol{x} = \lambda(c\boldsymbol{x})\,. \tag{4.28}$$

Thus, all vectors that are collinear to $\boldsymbol{x}$ are also eigenvectors of $\boldsymbol{A}$.

♢

Theorem 4.8 *$\lambda \in \mathbb{R}$ is eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ if and only if λ is a root of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ of $\boldsymbol{A}$.*

Definition 4.9 (Eigenspace and Eigenspectrum) For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the set of all eigenvectors of $\boldsymbol{A}$ associated with an eigenvalue λ spans a subspace of $\mathbb{R}^n$, which is called the *eigenspace* of $\boldsymbol{A}$ with respect to λ and is denoted by E_λ. The set of all eigenvalues of $\boldsymbol{A}$ is called the *eigenspectrum*, or just spectrum, of $\boldsymbol{A}$.

eigenspace

eigenspectrum

There are a number of ways to think about these characteristics:

- The eigenvector is a special vector that, left multiplying with the matrix $\boldsymbol{A}$ merely stretches the vector by a factor – the eigenvalue.
- Recall the definition of the kernel from Section 2.7.3, it follows that $E_\lambda = \ker(\boldsymbol{A} - \lambda\boldsymbol{I})$ since

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \iff \boldsymbol{A}\boldsymbol{x} - \lambda\boldsymbol{x} = \boldsymbol{0} \tag{4.29}$$

$$\iff (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \iff \boldsymbol{x} \in \ker(\boldsymbol{A} - \lambda\boldsymbol{I}). \tag{4.30}$$

- Similar matrices (see Definition 2.21) possess the same eigenvalues. Therefore, a linear mapping Φ has eigenvalues that are independent from the choice of basis of its transformation matrix. This makes eigenvalues, together with the determinant and the trace, key characteristic parameters of a linear mapping as they are all invariant under basis change.

Example 4.4 (Eigenvalues, Eigenvectors and Eigenspaces)
Here is an example of how to find the eigenvalues and eigenvectors of a 2×2 matrix.

$$\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}. \tag{4.31}$$

Step 1: Characteristic Polynomial

From our definition of the eigenvector $\boldsymbol{x}$ and eigenvalue λ for $\boldsymbol{A}$ there will be a vector such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, i.e., $(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Because we want $\boldsymbol{x} \neq \boldsymbol{0}$ (so that the definition of the eigenvectors is not trivial), this requires that the kernel (nullspace) of $\boldsymbol{A} - \lambda\boldsymbol{I}$ contains more elements than just $\boldsymbol{0}$. This means that $\boldsymbol{A} - \lambda\boldsymbol{I}$ is not invertible and therefore $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$. Hence we need to compute the roots of the characteristic polynomial (Equation (4.23)).

Step 2: Eigenvalues

The characteristic polynomial is given as

$$p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \tag{4.32}$$

$$= \det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} \tag{4.33}$$

$$= (4-\lambda)(3-\lambda) - 2\cdot 1\,. \tag{4.34}$$

We factorize the characteristic polynomial

$$p(\lambda) = (4-\lambda)(3-\lambda) - 2 \cdot 1 = 10 - 7\lambda + \lambda^2 = (2-\lambda)(5-\lambda) \quad (4.35)$$

and obtain the roots $\lambda_1 = 2$ and $\lambda_2 = 5$.

Step 3: Eigenvectors and Eigenspaces

We find the eigenvectors that correspond to these eigenvalues by looking at vectors $\boldsymbol{x}$ such that

$$\begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} \boldsymbol{x} = \mathbf{0}\,. \quad (4.36)$$

For $\lambda = 5$ we obtain

$$\begin{bmatrix} 4-5 & 2 \\ 1 & 3-5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}\,. \quad (4.37)$$

We now solve this homogeneous equation system and obtain a solution space

$$E_5 = \text{span}[\begin{bmatrix} 2 \\ 1 \end{bmatrix}]\,, \quad (4.38)$$

where for $c \neq 0$ all vectors $c[2,1]^\top$ are eigenvectors for $\lambda = 5$. Note, that this eigenspace is one-dimensional (spanned by a single vector).

Analogously, we find the eigenvector for $\lambda = 2$ by solving the homogeneous equation system

$$\begin{bmatrix} 4-2 & 2 \\ 1 & 3-2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \mathbf{0}\,. \quad (4.39)$$

This means any vector $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ where $x_2 = -x_1$, such as $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ is an eigenvector with eigenvalue 2. The corresponding eigenspace is given as

$$E_2 = \text{span}[\begin{bmatrix} 1 \\ -1 \end{bmatrix}]\,. \quad (4.40)$$

The two eigenspaces E_5 and E_2 in our previous Example 4.4 are one-dimensional as they are each spanned spanned by a single vector. But in other cases we may have multiple eigenvalues (see Definition 4.10) and the eigenspace may have more than one dimension.

Graphical Intuition in Two Dimensions

Let us gain some intuition for determinants, eigenvectors, eigenvalues and how linear maps affect space. Figure 4.4 depicts five transformation matrices and their impact on a square grid of points. The square grid of points is contained within a box of dimensions 2×2 with its centre at the origin.

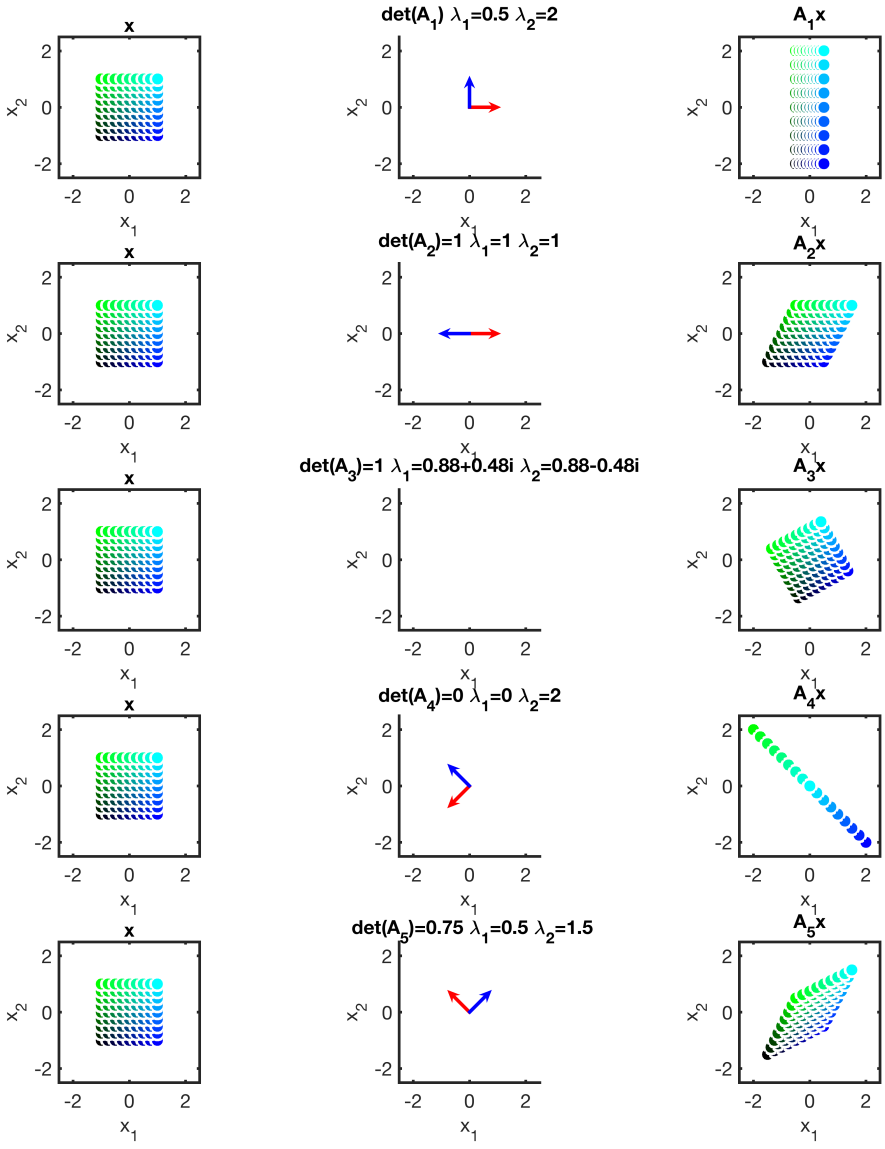


Figure 4.4 Determinants and eigenspaces. Overview of five linear mappings and their associated transformation matrices $\boldsymbol{A}_i \in \mathbb{R}^{2\times 2}$ project 81 color-coded points $\boldsymbol{x} \in \mathbb{R}^2$ (left column of plots) to target points $\boldsymbol{A}_i\boldsymbol{x}$ (right column of plots). The central column depicts the first eigenvector associated with eigenvalue λ_1, the second eigenvector associated with eigenvalue λ_2, as well as the value of the determinant. Each row depicts the effect of one of five transformation mappings in the standard basis $\boldsymbol{A}_i, i = \{1, \ldots, 5\}$.

- $\boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$. The direction of the two eigenvectors correspond to the canonical basis vectors in $\mathbb{R}^2$, i.e. to two cardinal axes. The horizontal axis is compressed by factor $\frac{1}{2}$ (eigenvalue $\lambda_1 = \frac{1}{2}$) and the vertical axis is extended by a factor of 2 (eigenvalue $\lambda_2 = 2$). The mapping is area preserving ($\det(\boldsymbol{A}_1) = 1 = 2 \times \frac{1}{2}$). Note, that while the area covered by the box of points remained the same, the circumference around the box has increased by 20%.
- $\boldsymbol{A}_2 = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$ corresponds to a shearing mapping , i.e., it shears the

points along the horizontal axis to the right if they are on the positive half of the vertical axis, and to the left vice versa. This mapping is area preserving ($\det(\boldsymbol{A}_2) = 1$). The eigenvalue $\lambda_1 = 1 = \lambda_2$ is repeated and hence the eigenvectors are co-linear (drawn here for emphasis in two opposite directions). This indicates that the mapping acts only along one direction (the horizontal axis). In geometry, the area preserving properties of this type of shearing parallel to an axis is also known as Cavalieri's principle of equal areas for parallelograms (Katz, 2004). Note, that the repeated identical eigenvalues make the two eigenvectors collinear, these are drawn in opposite directions to emphasize the shearing. Note, that the while the mapping is area preserving the circumference around the box of points has increased.

- $\boldsymbol{A}_3 = \begin{bmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$ The rotation matrix $\boldsymbol{A}_3$ rotates the points by $\frac{\pi}{6}$ (or 30° degrees) anti-clockwise, and has complex eigenvalues (reflecting that the mapping is a rotation) and no real valued eigenvalues (hence no eigenvectors are drawn). A pure rotation has to be area preserving, and hence the determinant is 1. Moreover, the circumference around the box of points has not changed. For more details on rotations we refer to Figure 3.15 in the corresponding section on rotations.
- $\boldsymbol{A}_4 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ reflects a mapping in the standard basis that collapses a two-dimensional domain onto a one-dimensional image space, hence the area is 0. We can see this because one eigenvalue is 0, collapsing the space in direction of the (red) eigenvector corresponding to $\lambda_1 = 0$, while the orthogonal (blue) eigenvector stretches space by a factor of $2 = \lambda_2$. Note, that while the area of the box of points vanishes the circumference does increase by around 41%.
- $\boldsymbol{A}_5 = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ is a shear-and-stretch mapping that shrinks space space by 75% ($|\det(\boldsymbol{A}_5)| = \frac{3}{4}$), stretching space along the (blue) eigenvector of λ_2 by 50% and compressing it along the orthogonal (red) eigenvector by a factor of 50%.

Remark The following statements are equivalent:

- λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$
- There exists an $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ with $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ or equivalently, $(\boldsymbol{A} - \lambda\boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$ can be solved non-trivially, i.e., $\boldsymbol{x} \neq \boldsymbol{0}$.
- $\mathrm{rk}(\boldsymbol{A} - \lambda\boldsymbol{I}_n) < n$
- $\det(\boldsymbol{A} - \lambda\boldsymbol{I}_n) = 0$

♢

Example 4.5 The curious case of the Identity matrix
The identity matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ has characteristic polynomial $p_{\boldsymbol{I}}(\lambda) = \det(\boldsymbol{I} - \lambda) = (1-\lambda)^n = 0$, which has only one eigenvalue $\lambda = 1$ that occurs n times. Moreover, we because $\boldsymbol{I}\boldsymbol{x} = \lambda\boldsymbol{x} = 1\boldsymbol{x}$ holds for all vectors $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. Because of this the sole Eigenspace E_1 of the identity matrix spans n dimensions, as e.g. all n standard basis vectors of $\mathbb{R}^n$ are eigenvectors of $\boldsymbol{I}$. We will revisit the concept of eigenvalue multiplicity further down we will revisit in Definition 4.10.

Remark (Eigenvalues and Eigenspaces) If λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ then the corresponding eigenspace E_λ is the solution space of the homogeneous linear equation system $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Geometrically, the eigenvector corresponding to a nonzero eigenvalue points in a direction that is stretched by the linear mapping, and the eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction is of the stretching is flipped. In particular, the eigenvector flips direction but remains collinear. ♢

Useful properties regarding eigenvalues and eigenvectors of various matrix types include

- A matrix $\boldsymbol{A}$ and its transpose $\boldsymbol{A}^\top$ possess the same eigenvalues, but not necessarily the same eigenvectors.
- Symmetric matrices always have real-valued eigenvalues.
- Symmetric positive definite matrices always have positive, real eigenvalues.
- The eigenvectors of symmetric matrices are always orthogonal to each other.

Example 4.6 (Eigenspectrum of a biological neural network)
Methods to analyze and learn from network data are an essential component of machine learning methods. The key to understanding networks is the connectivity between network nodes, especially if two nodes are connected to each other or not. In data science applications, it is often useful to study the matrix that captures this connectivity data. In Figure 4.5, we see the left plot showing the connectivity matrix (277×277), also referred to as adjacency matrix, of the complete neural network of the worm *C. Elegans*. Each row/column represents one of the 277 neurons of this worm's brain and the connectivity matrix $\boldsymbol{A}$ has a value of $a_{ij} = 1$ (white pixel) if neuron i talks to neuron j through a synapse, and $a_{ij} = 0$ (black pixel) otherwise. The neural network connectivity matrix is not symmetric, which implies that eigenvalues may not be real valued. Therefore we compute a version of the connectivity matrix as follows $\boldsymbol{A}_{sym} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^\top)$. This new matrix $\boldsymbol{A}_{sym}$ has a value of 1 whenever

two neurons are connected (irrespective of the direction of the connection) and zero otherwise. In the right panel, we show the eigenspectrum of $\boldsymbol{A}_{sym}$ in a scatter plot, on the horizontal axis we have the order of the eigenvalues from the largest (left most) to smallest eigenvalue and on the vertical axis the absolute of the eigenvalue. The S-like shape of this eigenspectrum is typical for many biological neural networks, the underlying mechanism responsible for this is an area of active neuroscience research.

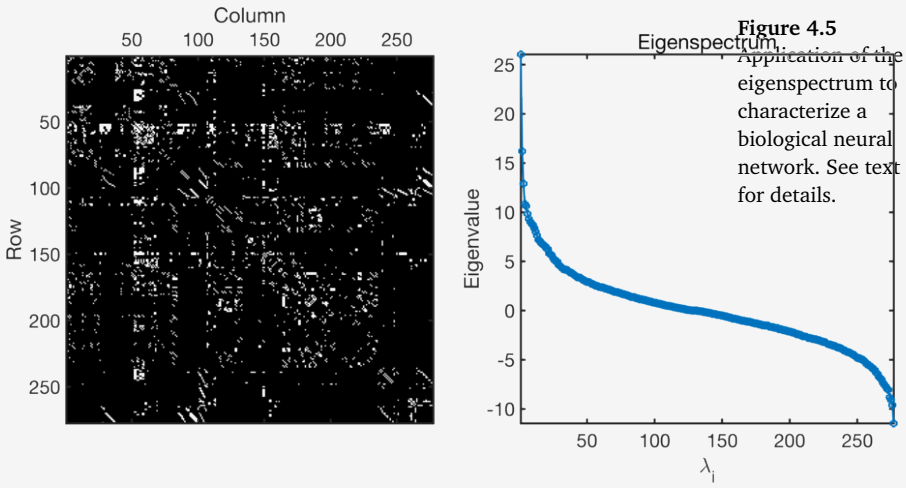


Figure 4.5 Application of the eigenspectrum to characterize a biological neural network. See text for details.

Definition 4.10 Let a square matrix $\boldsymbol{A}$ have an eigenvalue λ_i. The *algebraic multiplicity* of λ_i is the number of times the root appears in the characteristic polynomial. algebraic multiplicity

Definition 4.11 Let a square matrix $\boldsymbol{A}$ have an eigenvalue λ_i. The *geometric multiplicity* of λ_i is the total number of linearly independent eigenvectors associated with λ_i. In other words it is the dimensionality of the eigenspace spanned by the eigenvectors associated with λ_i. geometric multiplicity

Remark A specific eigenvalue's geometric multiplicity must be at least one, as by definition every eigenvalue has at least one associated eigenvector. An eigenvalue's geometric multiplicity cannot exceed its algebraic multiplicity, but it may be lower. ♢

Example 4.7

The matrix $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$ has two repeated eigenvalues $\lambda_1 = \lambda_2 = 2$ and an algebraic multiplicity of 2. The eigenvalue has however only one distinct eigenvector $\boldsymbol{x}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and thus geometric multiplicity 1.

defective matrix

Definition 4.12 A square matrix is a *defective matrix* if it does not have a complete set of eigenvectors (i.e. n linearly independent eigenvectors or the sum of the dimensions of the eigenspaces is n, see also Theorem 4.13) and is therefore not diagonalizable (see Section).

Remark Any defective matrix must have fewer than n distinct eigenvalues because distinct eigenvalues have linearly independent eigenvectors. Specifically, a defective matrix has at least one eigenvalue λ with an algebraic multiplicity $m > 1$ and fewer than m linearly independent eigenvectors associated with λ. ♢

Theorem 4.13 (Hogben (2013)) *Consider a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Then the eigenvectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are linearly independent.*

This theorem states that eigenvectors belonging to different eigenvalues form a linearly independent set. For symmetric matrices we can state a stronger version of Theorem 4.13.

Theorem 4.14 *Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ we can always obtain an $\boldsymbol{S}$ that is a symmetric positive semi-definite matrix by computing*

$$\boldsymbol{S} = \boldsymbol{A}^\top \boldsymbol{A}\,. \tag{4.41}$$

Understanding why this theorem holds is insightful for how we can use symmetrised matrices: Symmetry requires $\boldsymbol{S} = \boldsymbol{S}^\top$ and by inserting (4.41) we obtain $\boldsymbol{S} = \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{A}^\top(\boldsymbol{A}^\top)^\top = (\boldsymbol{A}^\top\boldsymbol{A})^\top = \boldsymbol{S}^\top$. Moreover, positive semi-definiteness (Section 3.2.3) requires that $\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} \geqslant 0$ and inserting (4.41) we obtain $\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{x}^\top\boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{A}\boldsymbol{x}) \geqslant 0$, because the scalar product computes a sum of squares (which are themselves always positive or zero).

Theorem 4.15 (Meyer (2000)) *Any symmetric matrix $\boldsymbol{A} = \boldsymbol{A}^\top \in \mathbb{R}^{n\times n}$ has n independent eigenvectors that form an orthogonal basis for $\mathbb{R}^n$.*

We should unpack this theorem as there are two implicit cases . First, if $\boldsymbol{A}$ is not defective than the situation is straightforward. Second, should $\boldsymbol{A}$ be symmetric and defective and thus have repeated identical eigenvalues, then the eigenvectors corresponding to the same eigenvalue do not have be orthogonal to each other. Fortunately, while beyond the scope of this textbook, we can always find a basis for the respective eigenspace made up of orthogonal eigenvectors: we take any basis for the eigenspace, and then apply the Gram-Schmidt process (see also Definition 3.9) to make it orthogonal.

Before we conclude our considerations of eigenvalues and eigenvectors it is useful to tie these matrix characteristics together with the previously covered concept of the determinant and the trace.

Theorem 4.16 *The determinant of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the product of its eigenvalues, i.e.,*

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i \,, \tag{4.42}$$

where λ_i are (possibly repeated) eigenvalues of $\boldsymbol{A}$.

Theorem 4.17 *The trace of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the sum of its eigenvalues, i.e.,*

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i \,, \tag{4.43}$$

where λ_i are (possibly repeated) eigenvalues of $\boldsymbol{A}$.

While we leave these two theorems without a proof, we point to the application of the determinant and trace of the characteristic polynomial as a way to derive them.

Remark A geometric intuition for these two theorems goes as follows (see also Figure 4.2 and corresponding text for other examples): Imagine a unit cube (a box with equal sides of length 1) in $\mathbb{R}^3$. We then map the 8 corner points of this box through our matrix $\boldsymbol{A}$ and obtain a new box, defined by the mapped 8 new corner points. We know that the eigenvalues capture the scaling of the basis with respect to the standard basis. Thus, they capture how the volume of the unit cube (which has volume 1) was transformed into our box. Thus, the determinant as product of eigenvalues is akin to the volume of the box, a large determinant suggests a large expansion of volume and vice versa. In contrast the trace is a sum of eigenvalues, i.e. a sum of length scales. Consider a gift ribbon we would want to tie around the box. The length of ribbon is proportional to the length of the sides of the box. The trace of $\boldsymbol{A}$ captures therefore a notion of how the matrix acts on the circumference of a volume. ♢

Example 4.8 (Google's PageRank – Webpages as Eigenvectors)
Google uses the eigenvector corresponding to the maximal eigenvalue of a matrix $\boldsymbol{A}$ to determine the rank of a page for search. The idea for the PageRank algorithm, developed at Stanford University by Larry Page and Sergey Brin in 1996, was that the importance of any web page can be approximated by the importance of pages that link to it. For this, they write down all websites as a huge directed graph that shows which page links to which. PageRank computes the weight (importance) $x_i \geqslant 0$ of a website a_i by counting the number of pages pointing to a_i. Moreover, PageRank takes into account the importance of the websites that link to a_i. The navigation behaviour of a user is then modelled by a transition matrix $\boldsymbol{A}$ of this graph that tells us with what (click) probability somebody will end

up on a different website. The matrix $\boldsymbol{A}$ has the property that for any initial rank/importance vector $\boldsymbol{x}$ of a website the sequence $\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}^2\boldsymbol{x}, \ldots$ converges to a vector $\boldsymbol{x}^*$. This vector is called the *PageRank* and satisfies $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{x}^*$, i.e., it is an eigenvector (with corresponding eigenvalue 1) of $\boldsymbol{A}$. After normalizing by $\boldsymbol{x}^*$, such that $\|\boldsymbol{x}^*\| = 1$ we can interpret the entries as probabilities. More details and different perspectives on PageRank can be found in the original technical report (Page et al., 1999).

PageRank

4.3 Cholesky Decomposition

There are many ways to factorize special types of matrices that we encounter often in machine learning. In the positive real numbers we have the square-root operation that yields us a decomposition of the number into components, for example, $9 = 3 \cdot 3$. For matrices, we need to be careful that we compute a square-root like operation on positive quantities. For symmetric, positive definite matrices (see Section 3.2.3) we can choose from a number of square-root equivalent operations. The *Cholesky decomposition* or *Cholesky factorization* provides a square-root equivalent operations that is very useful.

Cholesky decomposition

Cholesky factorization

Theorem 4.18 *Cholesky Decomposition: A symmetric positive definite matrix $\boldsymbol{A}$ can be factorized into a product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements:*

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix}. \tag{4.44}$$

$\boldsymbol{L}$ is called the Cholesky factor of $\boldsymbol{A}$.

Cholesky factor

Example 4.9

It is not immediately apparent why the Cholesky decomposition should exist for any symmetric, positive definite matrix. While we omit the proof we can go through an 3×3 matrix example.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \equiv \boldsymbol{L}\boldsymbol{L}^\top = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} \tag{4.45}$$

Expanding the right hand side yields

$$\boldsymbol{A} = \begin{bmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.$$

Comparing the left hand side and the right hand side shows that there is a simple pattern in the diagonal elements (l_{ii}):

$$l_{11} = \sqrt{a_{11}}\,, \qquad l_{22} = \sqrt{a_{22} - l_{21}^2}\,, \qquad l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)}\,. \tag{4.46}$$

Similarly for the elements below the diagonal (l_{ij}, where $i > j$) there is also a repeating pattern:

$$l_{21} = \frac{1}{l_{11}} a_{21}\,, \qquad l_{31} = \frac{1}{l_{11}} a_{31}\,, \qquad l_{32} = \frac{1}{l_{22}}(a_{32} - l_{31} l_{21})\,. \tag{4.47}$$

Thus, we have now constructed the Cholesky decomposition for any semi-positive definite 3×3 matrix. The key realization is that we can backwards calculate what the components l_{ij} for the $\boldsymbol{L}$ should be, given the values a_{ij} for $\boldsymbol{A}$ and previously computed values of l_{ij}.

The Cholesky decomposition is an important tool for the numerical computations underlying machine learning. Here, symmetric positive definite (SPD) matrices require frequent manipulation, for example, the covariance matrix $\boldsymbol{\Sigma}$ of a multivariate Gaussian variable (see Section 6.5) is SPD. We can apply the Cholesky decomposition to efficiently compute its inverse rather than directly solving for the inverse. Thus, under the hood of many numerical machine learning software packages the Cholesky decomposition is making computations more efficient.

4.4 Eigendecomposition and Diagonalization

Diagonal matrices are of the form

$$\boldsymbol{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix} \tag{4.48}$$

and possess a very simple structure. Therefore, they allow fast computation of determinants, powers and inverses. The determinant is the product of its diagonal entries, a matrix power $\boldsymbol{D}^k$ is given by each diagonal element raised to the power k, and the inverse $\boldsymbol{D}^{-1}$ is the reciprocal of its diagonal elements if all of them are non-zero.

In this section, we will look at how to transform matrices into diagonal form. This is an important application of the basis change we discussed in Section 2.7.2 and eigenvalues from Section 4.2.

Let us recall that two matrices $\boldsymbol{A}, \boldsymbol{D}$ are similar (Definition 2.21) if there exists an invertible matrix $\boldsymbol{P}$, such that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$. More specifically, we will look at matrices $\boldsymbol{A}$ that are similar to a diagonal matrix $\boldsymbol{D}$ that contains the eigenvalues of $\boldsymbol{A}$ on its diagonal.

diagonalizable **Definition 4.19** (Diagonalizable) A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *diagonalizable* if it is similar to a diagonal matrix, in other words, if there exists a matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ so that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$.

In the following, we will see that diagonalizing a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a way of expressing the same linear mapping but in another basis (see Section 2.6.1). Specifically we will try to diagonalize a matrix $\boldsymbol{A}$ by finding a new basis that consists of the eigenvectors of $\boldsymbol{A}$. We present two theorems, first for square matrices (Theorem 4.20) then for symmetric matrices (Theorem 4.21). The following results parallels the discussion we had about eigenvalues and eigenvectors (Theorem 4.13 and Theorem 4.15).

We first explore how to compute $\boldsymbol{P}$ so as to diagonalize $\boldsymbol{A}$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, let $\lambda_1, \dots, \lambda_n$ be a set of scalars, and let $\boldsymbol{p}_1, \dots, \boldsymbol{p}_n$ be a set of vectors in $\mathbb{R}^n$. Then we set $\boldsymbol{P} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_n]$ and let $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_n$. Then we can show that

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D} \tag{4.49}$$

if and only if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\boldsymbol{A}$ and the $\boldsymbol{p}_i$ are corresponding eigenvectors of $\boldsymbol{A}$.

We can see that this statement holds because

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}[\boldsymbol{p}_1, \dots, \boldsymbol{p}_n] = [\boldsymbol{A}\boldsymbol{p}_1, \dots, \boldsymbol{A}\boldsymbol{p}_n] \tag{4.50}$$

$$\boldsymbol{P}\boldsymbol{D} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_n]\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix} = [\lambda_1\boldsymbol{p}_1, \dots, \lambda_n\boldsymbol{p}_n] \quad . \tag{4.51}$$

Thus, (4.49) implies that

$$\boldsymbol{A}\boldsymbol{p}_1 = \lambda_1\boldsymbol{p}_1 \tag{4.52}$$

$$\vdots$$

$$\boldsymbol{A}\boldsymbol{p}_n = \lambda_n\boldsymbol{p}_n \tag{4.53}$$

and vice versa.

Thus, the matrix $\boldsymbol{P}$ must be composed of columns of eigenvectors. But this is not sufficient to know if we can diagonalize $\boldsymbol{A}$, as our definition of diagonalization requires that $\boldsymbol{P}$ is invertible. From Theorem 4.3 we know that our square matrix $\boldsymbol{P}$ is only invertible (has determinant $\neq 0$) if it has full rank. This implies that the eigenvectors $\boldsymbol{p}_1, \dots, \boldsymbol{p}_n$ must be linearly independent. Moreover, consider that Theorem 4.13 tells us that when $\boldsymbol{A}$ has n independent eigenvectors, then it also has n distinct eigenvalues. Taking together these arguments we can now combine them to formulate a key theorem of this chapter.

Diagonalization **Theorem 4.20** *Eigendecomposition/Diagonalization theorem. A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1} \tag{4.54}$$

where $\boldsymbol{P}$ is an invertible matrix of eigenvectors and $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{A}$, if and only if $\boldsymbol{A}$ has n independent eigenvectors (i.e. $\mathrm{rk}(\boldsymbol{P}) = n$).

Remark The *Jordan Normal Form* of a matrix offers a decomposition that works for defective matrices but is beyond the scope of this book (Lang, 1987). ♢

Jordan Normal Form

For symmetric matrices we can obtain even stronger outcomes for the eigenvalue decomposition.

Theorem 4.21 *A symmetric matrix $\boldsymbol{S} = \boldsymbol{S}^\top \in \mathbb{R}^{n\times n}$ can always be diagonalized into*

$$\boldsymbol{S} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top \tag{4.55}$$

where $\boldsymbol{P}$ is matrix of n orthonormal vectors and $\boldsymbol{D}$ is a diagonal matrix of its n eigenvalues.

Remark Should the matrix $\boldsymbol{S}$ not be defective and thus have n distinct eigenvalues, then the n vectors making up $\boldsymbol{P}$ will be the orthogonal eigenvectors of $\boldsymbol{S}$. However, should $\boldsymbol{S}$ be defective and thus with repeated identical eigenvalues, then the eigenvectors can be made orthogonal to each other (see discussion of Theorem 4.15). ♢

Proof From Theorem 4.15 we know that $\boldsymbol{P} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_n]$ has n orthogonal eigenvectors of $\boldsymbol{S}$ with eigenvalues $\lambda_1, \dots, \lambda_n$. We can then write

$$(\boldsymbol{P}^\top \boldsymbol{P})_{ij} = \boldsymbol{p}_i^\top \boldsymbol{p}_j \tag{4.56}$$

where

$$\boldsymbol{p}_i^\top \boldsymbol{p}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.57}$$

and therefore $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}$ and $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$.

We observe the following product

$$\lambda_i \boldsymbol{P}^\top \boldsymbol{p}_i = \lambda_i \left[\boldsymbol{p}_1, \dots, \boldsymbol{p}_n\right]^\top \boldsymbol{p}_i = \lambda_i \boldsymbol{e}_i, \tag{4.58}$$

which we will use in the following derivation.

$$\begin{aligned}
\boldsymbol{P}^\top \boldsymbol{S}\boldsymbol{P} &= \boldsymbol{P}^\top \boldsymbol{S}\left[\boldsymbol{p}_1, \dots, \boldsymbol{p}_n\right] && (4.59)\\
&= \boldsymbol{P}^\top \left[\boldsymbol{S}\boldsymbol{p}_1, \dots, \boldsymbol{S}\boldsymbol{p}_n\right] && (4.60)\\
&= \boldsymbol{P}^\top \left[\lambda_1 \boldsymbol{p}_1, \dots, \lambda_n \boldsymbol{p}_n\right] && (4.61)\\
&= \left[\boldsymbol{p}_1, \dots, \boldsymbol{p}_n\right]^\top \left[\lambda_1 \boldsymbol{p}_1, \dots, \lambda_n \boldsymbol{p}_n\right] && (4.62)\\
&= \left[\lambda_1 \boldsymbol{e}_1, \dots, \lambda_n \boldsymbol{e}_n\right] = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = \boldsymbol{D} && (4.63)
\end{aligned}$$

□

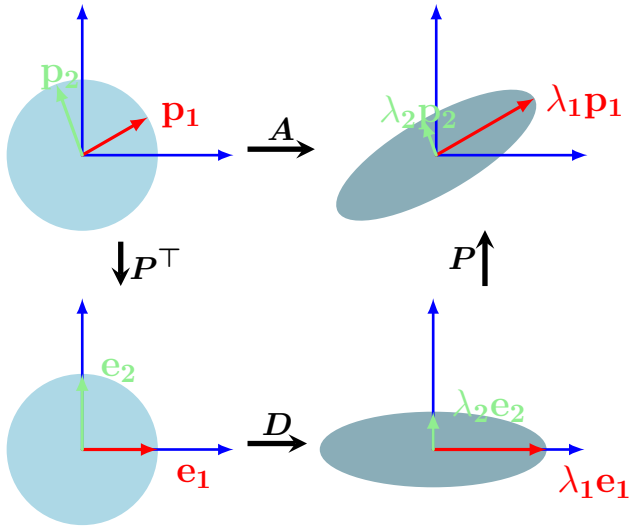


Figure 4.6 Intuition behind the eigendecomposition of a $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ in the standard basis as sequential transformations. Top-left to bottom-left: $\boldsymbol{P}^\top$ performs a basis change (here drawn in $\mathbb{R}^2$ and depicted as a rotation-like operation) mapping the eigenvectors into the standard basis. Bottom-left-to-bottom right $\boldsymbol{D}$ performs a scaling along the remapped orthogonal eigenvectors, depicted here by a circle being stretched to an ellipse. Bottom-right to top-right: $\boldsymbol{P}$ undoes the basis change (depicted as a reverse rotation) and restores the original coordinate frame.

Geometric intuition for the eigendecomposition

We can interpret the eigendecomposition of a matrix as follows (see also Figure 4.6): Let $\boldsymbol{A}$ be the transformation matrix of a linear mapping with respect to the standard basis. $\boldsymbol{P}^{-1}$ performs a basis change from the standard basis into the eigenbasis. This maps the eigenvectors $\boldsymbol{p}_i$ (red and green arrows in Figure 4.6) onto the standard axes $\boldsymbol{e}_i$. Then, the diagonal $\boldsymbol{D}$ scales the vectors along these axes by the eigenvalues $\lambda_i \boldsymbol{e}_i$ and, finally, $\boldsymbol{P}$ transforms these scaled vectors back into the standard/canonical coordinates (yielding $\lambda_i \boldsymbol{p}_i$).

Example 4.10

Let us compute the eigendecomposition of a (symmetric) matrix $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

Step 1: Compute the eigenvalues and eigenvectors

The matrix has eigenvalues

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I}) &= \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\right) && (4.64) \\ &= (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 \\ &= (\lambda - 3)(\lambda - 1) = 0. && (4.65) \end{aligned}$$

So the eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = 1$ and $\lambda_2 = 3$ and the associated normalized eigenvectors are obtained via

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \boldsymbol{p}_1 = 1\boldsymbol{p}_1 \qquad (4.66)$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \boldsymbol{p}_2 = 3\boldsymbol{p}_2 \,. \tag{4.67}$$

This yields

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} , \boldsymbol{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} . \tag{4.68}$$

Step 2: Check for existence

The matrix is symmetric, we therefore know that the eigenvectors are linearly independent and the eigenvalues are distinct (but we can also quickly eye-ball this to validate our calculations), and so a diagonalization is possible.

Step 3: Construct the matrix $\boldsymbol{P}$ to diagonalize $\boldsymbol{A}$

To compute the diagonalizing matrix we collect these normalized eigenvectors together

$$\boldsymbol{P} = [\boldsymbol{p}_1, \, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \tag{4.69}$$

so that we obtain

$$\begin{aligned} \boldsymbol{A}\boldsymbol{P} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 3 \\ -1 & 3 \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} = \boldsymbol{P}\boldsymbol{D} \,. \end{aligned} \tag{4.70}$$

We can now obtain the matrices of the eigendecomposition by right multiplying with $\boldsymbol{P}^{-1}$. Alternatively as the matrix $\boldsymbol{A}$ is symmetric we can use the orthogonality property of its eigenvectors with $\boldsymbol{P}^\top = \boldsymbol{P}^{-1}$ and solve for $\boldsymbol{A}$ directly to obtain the eigendecomposition:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top \tag{4.71}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} . \tag{4.72}$$

The eigenvalue decomposition of a matrix has a number of convenient properties

- Diagonal matrices $\boldsymbol{D}$ have the nice property that they can be efficiently raised to a power. Therefore we can find a matrix power for a general matrix $\boldsymbol{A}$ via the eigenvalue decomposition

$$\boldsymbol{A}^k = (\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1})^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1} \,. \tag{4.73}$$

  Computing $\boldsymbol{D}^k$ is efficient because we apply this operation individually to any diagonal element.

- A different property of diagonal matrices is that they can be used to decouple variables. This will be important in probability theory to interpret random variables, e.g., for the Gaussian distributions we will encounter in Section 6.5 and in applications such as dimensionality reduction Chapter 10.

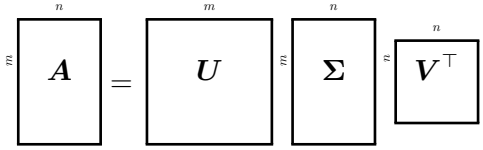
The eigenvalue decomposition requires square matrices, and for non-symmetric square matrices it is not guaranteed that we can transform them into diagonal form. It would be useful to be able to perform a decomposition on general matrices. In the next section, we introduce a more general matrix decomposition technique, the Singular Value Decomposition.

4.5 Singular Value Decomposition

The Singular Value Decomposition (SVD) of a matrix is a central matrix decomposition method in linear algebra. It has been referred to as the "fundamental theorem of linear algebra" (Strang, 1993) because it can be applied to all matrices, not only to square matrices, and it always exists. Moreover, as we will explore in the following, the SVD of a linear mapping $\Phi : V \to W$ quantifies the resulting change between the underlying geometry of these two vector spaces. We recommend Kalman (1996); Roy and Banerjee (2014) for a deeper overview of the mathematics of the SVD.

Theorem 4.22 (SVD theorem) *Let $\boldsymbol{A}^{m\times n}$ be a rectangular matrix of rank r, with $r \in [0, \min(m, n)]$. The Singular Value Decomposition or SVD of $\boldsymbol{A}$ is a decomposition of $\boldsymbol{A}$ of the form*

Singular Value Decomposition
SVD



(4.74)

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix composed of column vectors $\boldsymbol{u}_i$, $i = 1, \ldots, m$, and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix of column vectors $\boldsymbol{v}_j$, $j = 1, \ldots, n$, and $\boldsymbol{\Sigma}$ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geqslant 0$ and $\Sigma_{ij} = 0$, $i \neq j$. The SVD is always possible for any matrix $\boldsymbol{A}$.

singular values
left-singular vectors
right-singular vectors

The σ_i are called the singular values, $\boldsymbol{u}_i$ are called the left-singular vectors and $\boldsymbol{v}_j$ are called the right-singular vectors. By convention the singular values are ordered, i.e., $\sigma_1 \geqslant \sigma_2 \geqslant \sigma_r \geqslant 0$ and correspondingly so are the singular vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$.

We will see a proof of this theorem later in this section. The SVD allows us to decompose general matrices, and the existence of the unique *singular value matrix* $\boldsymbol{\Sigma}$ requires attention. Observe that the $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is rectangular, that is it is non-square. In particular note that $\boldsymbol{\Sigma}$ is the same size as $\boldsymbol{A}$. This means that $\boldsymbol{\Sigma}$ has a diagonal submatrix that contains the

singular value matrix

singular values and needs additional zero vectors that increase the dimension.

Specifically, if $m > n$ then the matrix $\mathbf{\Sigma}$ has diagonal structure up to row n and then consists of $\mathbf{0}^\top$ row vectors from $n+1$ to m below

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} . \tag{4.75}$$

Conversely, if $m < n$ the matrix $\mathbf{\Sigma}$ has a diagonal structure up to column m and columns that consist of $\mathbf{0}$ from $m+1$ to n.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & & 0 \\ 0 & 0 & \sigma_n & 0 & \dots & 0 \end{bmatrix} \tag{4.76}$$

4.5.1 Geometric Intuitions for the SVD

The SVD has a number of interesting geometric intuitions to offer to describe a transformation matrix. Broadly there are two intuitive views we can have. First we consider the SVD as sequential operations performed on the bases (discussed in the following), and second we consider the SVD as operations performed on sets of (data) points as described in Example 4.11.

The SVD can be interpreted as a decomposition of a linear mapping (recall Section 2.7.1) $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ into three operations (see Figure 4.7 for the following). The SVD intuition follows superficially a similar structure to our eigendecomposition intuition (confront Figure 4.7 for the SVD with Figure 4.6 for the eigendecomposition): Broadly speaking the SVD performs a basis change ($\boldsymbol{V}^\top$) followed by a scaling and augmentation (or reduction) in dimensionality ($\mathbf{\Sigma}$) and then performs a second basis change ($\boldsymbol{U}$). The SVD entails a number of important details and caveats which is why we will review our intuition in more detail and precision.

Assume we are given a transformation matrix of Φ with respect to the standard bases B and C of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Moreover, assume a second basis $\tilde{B}$ of $\mathbb{R}^n$ and $\tilde{C}$ of $\mathbb{R}^m$. Then

1 $\boldsymbol{V}$ performs a basis change in the domain $\mathbb{R}^n$ from $\tilde{B}$ (represented by the red and green vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ in Figure 4.7 top left) to the canonical basis B. It is useful here to recall our discussion of basis changes (Section 2.7.2), orthogonal matrices (Definition 3.8) and orthonormal bases (Section 3.5), as $\boldsymbol{V}^\top = \boldsymbol{V}^{-1}$ performs a basis change from B to

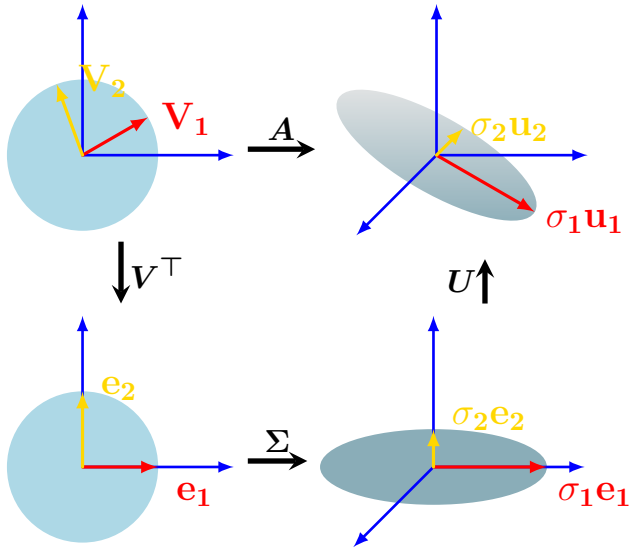


Figure 4.7 Intuition behind SVD of a $\boldsymbol{A} \in \mathbb{R}^{3\times 2}$ in the standard basis as sequential transformations. Top-left to bottom-left: $\boldsymbol{V}^\top$ performs a basis change in $\mathbb{R}^2$. Bottom-left-to-bottom right $\boldsymbol{\Sigma}$ performs a scaling and increases the dimensionality from $\mathbb{R}^2$ to $\mathbb{R}^3$. The ellipse in the bottom-right lives in $\mathbb{R}^3$ and the third dimension is orthogonal to the surface of the elliptical disk. Bottom-right to top-right: $\boldsymbol{U}$ performs a second basis change within $\mathbb{R}^3$.

$\tilde{B}$ (the red and green vectors are now aligned with the canonical basis in Figure 4.7 bottom left).

2 Having changed the coordinate system to $\tilde{B}$, $\boldsymbol{\Sigma}$ scales the new coordinates by the singular values σ_i (and adding or deleting dimensions), i.e., $\boldsymbol{\Sigma}$ is the transformation matrix of Φ with respect to $\tilde{B}$ and $\tilde{C}$ (represented by the red and green vectors being stretched and lying in the $\boldsymbol{e}_1$-$\boldsymbol{e}_2$ plane which is now embedded in a third dimension in Figure 4.7 bottom right) .

3 $\boldsymbol{U}$ performs a basis change in the codomain $\mathbb{R}^m$ from $\tilde{C}$ into the canonical basis of $\mathbb{R}^m$ (represented by a rotation of red and green vectors out of the plane of the $\boldsymbol{e}_1$-$\boldsymbol{e}_2$ plane in Figure 4.7 bottom right).

The SVD expresses a change of basis in both the domain and codomain: The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are the bases $\tilde{B}$ of $\mathbb{R}^n$ and $\tilde{C}$ of $\mathbb{R}^m$, respectively. Note, how this is in contrast with the eigendecomposition that operates within the same vector space (where the same basis change is applied and then undone). What makes the SVD special is that these two (different) bases are simultaneously linked by the singular values matrix $\boldsymbol{\Sigma}$. We refer to Section 2.7.2 and Figure 2.11 for a more detailed discussion on basis change.

Example 4.11
Data points and the SVD. Consider a mapping of a square grid of points $\mathcal{X} \in \mathbb{R}^2$ which fit in a box of size 2×2 centered at the origin. Using the

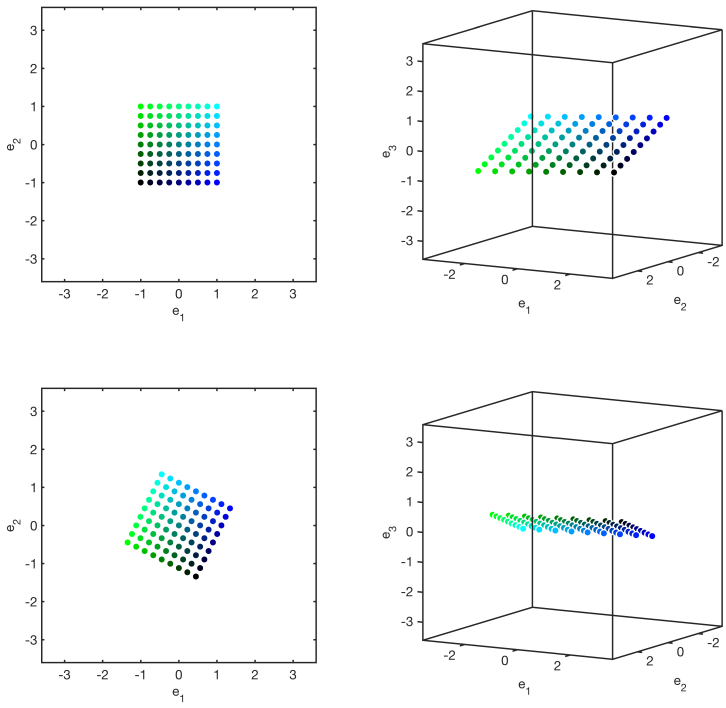


Figure 4.8 SVD and mapping of data points. The panels follow the same anti-clockwise structure of Figure 4.7. See main text for details.

standard basis we map these points using

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{4.77}$$

$$= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{4.78}$$

$$= \begin{bmatrix} 0.913 & 0 & -0.408 \\ -0.365 & 0.4472 & -0.816 \\ 0.182 & 0.894 & 0.4082 \end{bmatrix} \begin{bmatrix} 2.449 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4472 & -0.894 \\ 0.8941 & 0.4472 \end{bmatrix} \tag{4.79}$$

.

We start with a set of points $\mathcal{X}$ (colored dots, see top left panel of Figure 4.8) arranged in a grid.

The points $\mathcal{X}$ after rotating them using $\boldsymbol{V}^\top \in \mathbb{R}^{2\times 2}$ are shown in the bottom-left panel of Figure 4.8. After a mapping $\boldsymbol{\Sigma}$ to the codomain $\mathbb{R}^3$ (see bottom right panel in Figure 4.8) we can see how all the points lie on the $\boldsymbol{e}_1$-$\boldsymbol{e}_2$ plane. The third dimension was added, and the arrangement of points has been stretched by the singular values.

The direct mapping of the points $\mathcal{X}$ by $\boldsymbol{A}$ to the codomain $\mathbb{R}^3$ equals the transformation of $\mathcal{X}$ by $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ performs a rotation within the codomain $\mathbb{R}^3$ so that the mapped points are no longer restricted to the $\boldsymbol{e}_1$-$\boldsymbol{e}_2$ plane; they still are on a plane (see top-right panel of Figure 4.8).

4.5.2 Existence and Construction of the SVD

We will next discuss why the SVD exists and show how to compute it in detail. The SVD of a general matrix is related to the eigendecomposition of a square matrix and has some similarities.

Remark Compare the eigenvalue decomposition of a symmetric positive definite (SPD) matrix

$$\boldsymbol{S} = \boldsymbol{S}^\top = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top \tag{4.80}$$

(which always exists) to the structure of the SVD of

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top . \tag{4.81}$$

We identify

$$\boldsymbol{U} = \boldsymbol{P} = \boldsymbol{V} , \tag{4.82}$$
$$\boldsymbol{D} = \boldsymbol{\Sigma} , \tag{4.83}$$

so that the SVD of SPD matrices is their eigenvalue decomposition. ♢

In the following we will explore why Theorem 4.22 should hold and how it is constructed. Computing the SVD of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is equivalent to finding two sets of orthonormal bases $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ and $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ of the domain $\mathbb{R}^m$ and the codomain $\mathbb{R}^n$, respectively. From these ordered bases we will construct the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.

Our plan is to start with constructing the orthonormal set of right-singular vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$. We then construct the orthonormal set of left-singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathbb{R}^n$. Thereafter, we will link the two and require that the orthogonality of the $\boldsymbol{v}_i$ is preserved under the transformation of $\boldsymbol{A}$. This is important because we know the images $\boldsymbol{A}\boldsymbol{v}_i$ form a set of orthogonal vectors. We will then need to normalize these images by scalar factors, which will turn out to be the singular values, so that the images are also normalized in length.

Let us begin with constructing the right-singular vectors. We have previously learned that the eigenvalue decomposition is a method to construct an orthonormal basis, and it always exists for symmetric matrices by Theorem 4.21. Moreover, from Theorem 4.14 we can always construct a symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{n\times n}$ from any rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

Thus, we can always diagonalize $\boldsymbol{A}^\top\boldsymbol{A}$ and obtain

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top = \boldsymbol{P}\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}\boldsymbol{P}^\top. \tag{4.84}$$

Take note that the $\lambda_i \geqslant 0$ are the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$. Let us assume the SVD of $\boldsymbol{A}$ exists and inject (4.74) into (4.84).

$$\boldsymbol{A}^\top\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top. \tag{4.85}$$

where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices. Therefore, with $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ we obtain

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_n^2\end{bmatrix}\boldsymbol{V}^\top. \tag{4.86}$$

Comparing now (4.84) and (4.86) we identify

$$\boldsymbol{V} = \boldsymbol{P}, \tag{4.87}$$

$$\sigma_i^2 = \lambda_i. \tag{4.88}$$

Therefore, the eigenvectors $\boldsymbol{P}$ of $\boldsymbol{A}^\top\boldsymbol{A}$ are the right-singular vectors $\boldsymbol{V}$ of $\boldsymbol{A}$ (see (4.87)). They form an orthogonal basis because of Theorem 4.21, for the domain of the SVD. We can always normalize this orthogonal basis to obtain an orthonormal basis. Moreover, the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are the squared singular values of $\boldsymbol{\Sigma}$ (see (4.88)).

Let us now repeat this derivation but this time we will focus on obtaining the left singular vectors $\boldsymbol{U}$ instead of $\boldsymbol{V}$. Therefore we start again by computing the SVD of a symmetric matrix, this time $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m\times m}$ (instead of the above $\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{n\times n}$). We inject again (4.74) and obtain:

$$\boldsymbol{A}\boldsymbol{A}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \tag{4.89}$$

$$= \boldsymbol{U}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_m^2\end{bmatrix}\boldsymbol{U}^\top. \tag{4.90}$$

We can now obtain from the same arguments about symmetric matrices and their diagonalization, now applied to $\boldsymbol{A}\boldsymbol{A}^\top$, the orthonormal eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$. These are the left-singular vectors $\boldsymbol{U}$ and form an orthonormal basis set in the codomain of the SVD.

This leaves the question of the structure of the matrix $\boldsymbol{\Sigma}$. We need to show that regardless of $n > m$ or $n < m$, that $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same non-zero eigenvalues: Let us assume that λ is a non-zero eigenvalue of $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{x}$ is an eigenvector belonging to λ. Thus, the eigenvalue equation

$$(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{x} = \lambda\boldsymbol{x} \tag{4.91}$$

can be manipulated by left multiplying by $\boldsymbol{A}^\top$ and pulling on the right-hand side the scalar factor λ forward. This yields

$$\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{x} = \boldsymbol{A}^\top(\lambda\boldsymbol{x}) = \lambda(\boldsymbol{A}^\top\boldsymbol{x}) \tag{4.92}$$

and we can use matrix multiplication associativity to reorder the left-hand side factors

$$(\boldsymbol{A}^\top\boldsymbol{A})(\boldsymbol{A}^\top\boldsymbol{x}) = \lambda(\boldsymbol{A}^\top\boldsymbol{x})\,. \tag{4.93}$$

This is the eigenvalue equation for $\boldsymbol{A}^\top\boldsymbol{A}$. Therefore, λ is the same eigenvalue for $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$. Thus, both matrices have the same non-zero eigenvalues and the $\boldsymbol{\Sigma}$ matrices in the SVD for both cases have to be the same.

The last step in developing our reasoning is to link up all the parts we touched on so far. We have now an orthonormal set of right-singular vectors in $\boldsymbol{V}$. But, to finish construction of the SVD we link them to the orthonormal vectors $\boldsymbol{U}$. To reach this goal we use the fact the images of the $\boldsymbol{v}_i$ under $\boldsymbol{A}$ have to be orthonormal, too. Using the results from Section 3.4, we require that the inner product between $\boldsymbol{A}\boldsymbol{v}_i$ and $\boldsymbol{A}\boldsymbol{v}_j$ must be 0 for $i \neq j$. For any two orthogonal eigenvectors $\boldsymbol{v}_i, \boldsymbol{v}_j$, $i \neq j$ it holds that

$$(\boldsymbol{A}\boldsymbol{v}_i)^\top(\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_j = \boldsymbol{v}_i^\top(\lambda_j\boldsymbol{v}_j) = \lambda_j\boldsymbol{v}_i^\top\boldsymbol{v}_j = 0\,. \tag{4.94}$$

For the case $m > r$ this holds for all pairs $\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r$ the images are a basis of $\mathbb{R}^m$, while if any further vectors $\boldsymbol{A}\boldsymbol{v}_i, i > r$ exist, they must be in the nullspace of $\boldsymbol{A}$ (see remark after proof for the converse case).

To complete the SVD construction we need left-singular vectors that are ortho*normal*: we normalize the images of the right-singular vectors $\boldsymbol{A}\boldsymbol{v}_i$ and call them $\boldsymbol{u}_i$,

$$\boldsymbol{u}_i = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\|\boldsymbol{A}\boldsymbol{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i \tag{4.95}$$

where the last equality was obtained from (4.88) and from equation (4.90) showing us that the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ are such that $\sigma_i^2 = \lambda_i$.

Therefore, the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$, which we know are the right-singular vectors $\boldsymbol{v}_i$ and their normalized images under $\boldsymbol{A}$, the left singular vectors $\boldsymbol{u}_i$, form two self-consistent sets of orthonomal bases that are coupled by the singular value matrix $\boldsymbol{\Sigma}$.

singular value equation

Remark Let us rearrange (4.95) to obtain the *singular value equation*

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i\,, \quad i = 1, \ldots, r\,. \tag{4.96}$$

This equation closely resembles the eigenvalue equation (4.27), but the vectors on the left and the right-hand sides are not the same.

For $n > m$ (4.96) holds only for $i \leqslant m$ and (4.96) says nothing about the $\boldsymbol{u}_i$ for $i > m$, but we know by construction that they are orthonormal.

Conversely for $m > n$, then (4.96) holds only for $i \leqslant n$. For $i > n$ we have $\boldsymbol{A}\boldsymbol{v}_i = 0$ and we still know that the $\boldsymbol{v}_i$ form an orthonormal set. This means that the SVD also supplies an orthonormal basis of the kernel (or null space) of $\boldsymbol{A}$, the set of vectors $\boldsymbol{x}$ with $\boldsymbol{A}\boldsymbol{x} = 0$ (see Section 2.7.3).

Moreover, horizontally concatenating the $\boldsymbol{v}_i$ as the columns of $\boldsymbol{V}$ and $\boldsymbol{u}_i$ as the columns of $\boldsymbol{U}$ yields

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}\,. \tag{4.97}$$

where $\boldsymbol{\Sigma}$ has the same dimensions as $\boldsymbol{A}$ and a diagonal structure for rows $1, \ldots, r$. Hence, right-multiplying with $\boldsymbol{V}^\top = \boldsymbol{V}^{-1}$ yields $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, which is again our singular value decomposition of $\boldsymbol{A}$. ♢

Example 4.12

Let us find the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}. \tag{4.98}$$

Step 1: Compute the symmetrized matrix $\boldsymbol{A}\boldsymbol{A}^\top$

$$\boldsymbol{A}\boldsymbol{A}^\top = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 2 \\ 1 & 0 & 1 \end{bmatrix}. \tag{4.99}$$

Step 2: Compute the eigenvalue decomposition of $\boldsymbol{A}\boldsymbol{A}^\top$

We compute the singular values and right-singular vectors through the eigenvalue decomposition of $\boldsymbol{A}^\top\boldsymbol{A}$

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{4.100}$$

$$= \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{2}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{2}} & \frac{-2}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top. \tag{4.101}$$

Note, that due to our orthonormality requirement this implies that we chose the 3rd column of $\boldsymbol{P}$ so as to be orthogonal to the other two columns. As the singular values σ_i are the square root of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ we obtain them straight from $\boldsymbol{D}$. Note that because $\mathrm{rk}(\boldsymbol{A}) = 2$ there are only two non-zero singular values, $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$. The singular value matrix must be the same size as $\boldsymbol{A}$, hence,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.102}$$

We also have obtained already the right-singular vectors because

$$\boldsymbol{V} = \boldsymbol{P} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{2}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \end{bmatrix} . \tag{4.103}$$

Step 3: Compute the normalized image of the right-singular vectors

We now find the left singular-vectors by computing the image of the right-singular vectors under $\boldsymbol{A}$ and normalizing them by dividing them by their corresponding singular value.

$$\boldsymbol{u}_1 = \frac{1}{\sigma_1}\boldsymbol{A}\boldsymbol{v}_1 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} \frac{5}{\sqrt{30}} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{bmatrix} , \tag{4.104}$$

$$\boldsymbol{u}_2 = \frac{1}{\sigma_2}\boldsymbol{A}\boldsymbol{v}_2 = \frac{1}{1}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} \tag{4.105}$$

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2] = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix} . \tag{4.106}$$

Note that on a computer the approach illustrated here has poor numerical behaviour, and the SVD of $\boldsymbol{A}$ is normally computed without resorting to the eigenvalue decomposition of the .

4.5.3 Eigenvalue Decomposition vs Singular Value Decomposition

Let us consider the eigendecomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ and SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and review the core elements of the past sections.

- The SVD always exists for any matrix $\mathbb{R}^{m\times n}$. The eigendecomposition is only defined for square matrices $\mathbb{R}^{n\times n}$ and only exists if we can find a basis of eigenvectors (or n independent eigenvectors).
- The vectors in the eigendecomposition matrix $\boldsymbol{P}$ are not necessarily orthogonal, so the change of basis is not a simple rotation and scaling. On the other hand, the vectors in the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ in the SVD are orthonormal, so they do represent rotations (or possibly reflections).
- Both the eigendecomposition and the SVD are compositions of three linear mappings:
  1. Change of basis in the domain
  2. Independent scaling of each new basis vector and mapping from domain to co-domain
  3. Change of basis in the co-domain

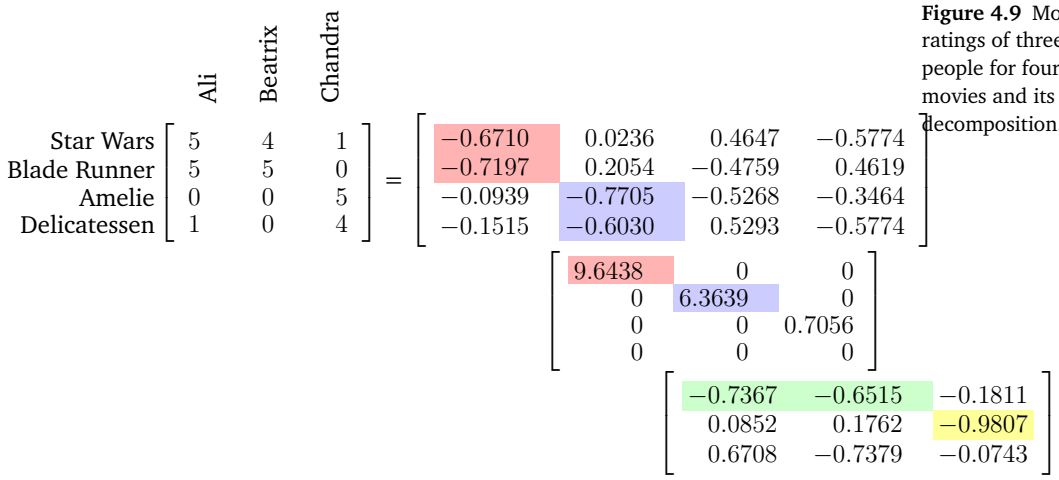


Figure 4.9 Movie ratings of three people for four movies and its SVD decomposition.

A key difference between the eigendecomposition and the SVD is that in the SVD, domain and co-domain can be vector spaces of different dimensions.

- In the SVD, the left and right singular vector matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are generally not inverse of each other. In the eigendecomposition the eigenvector matrices $\boldsymbol{P}$ and $\boldsymbol{P}^{-1}$ are inverses of each other.
- In the SVD, the entries in the diagonal matrix $\boldsymbol{\Sigma}$ are all real and nonnegative, which is not generally true for the diagonal matrix in the eigendecomposition.
- The SVD and the eigendecomposition are closely related through their projections
  - The left-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$
  - The right-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$.
  - The non-zero singular values of $\boldsymbol{A}$ are the square roots of the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$, and are equal to the non-zero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$.
- For symmetric matrices the eigenvalue decomposition and the SVD are one and the same.

Example 4.13 (Finding Structure in Movie Ratings and Consumers)
Let us understand a way to interpret the practical meaning of the SVD by analysing data on people and their preferred movies. Consider 3 viewers (Ali, Beatrix, Chandra) rating 4 different movies (Star Wars, Blade Runner, Amelie, Delicatessen). Their ratings are values between 0 (worst) and 5 (best) and encoded in a data matrix $\boldsymbol{A} \in \mathbb{R}^{4\times 3}$ (see Figure 4.9). Each row represents a movie and each column a user. Thus, the column vectors of movie ratings, one for each viewer, are $\boldsymbol{x}_{\text{Ali}}$, $\boldsymbol{x}_{\text{Beatrix}}$, $\boldsymbol{x}_{\text{Chandra}}$.

Factoring $\boldsymbol{A}$ using SVD provides a way to capture the relationships of how people rate movies, and especially if there is a structure linking which people like which movies. Applying the SVD to our data matrix makes a number of assumptions

1 All viewers rate movies consistently using the same linear mapping.
2 There are no errors or noise in the ratings data.
3 We interpret the left-singular vectors $\boldsymbol{u}_i$ as stereotypical movies and the right-singular vectors $\boldsymbol{v}_j$ as stereotypical viewers.

We then make the assumption that any viewer's specific movie preferences can be expressed as a linear combination of the $\boldsymbol{v}_j$s. Similarly, any movie's like-ability can be expressed as a linear combination of the $\boldsymbol{u}_i$s. Thus, a vector in the domain of the SVD can be interpreted as a viewer in the "space" of stereotypical viewers and a vector in the co-domain of the SVD correspondingly as a movie in the "space" of stereotypical movies (these two "spaces" are only meaningfully spanned by the respective viewer and movie data that we have, if the data itself covers sufficient diversity of viewers and movies). Let us look at the specific outcome of performing SVD: The first left-singular vector $\boldsymbol{u}_1$ has large absolute values for the two science fiction movies and a large first singular value (red shading in Figure 4.9). Thus, this groups a type of users with a set of movies – we interpret this here as the notion of a science fiction theme. Similarly, the first right-singular $\boldsymbol{v_1}$ shows large absolute values for Ali and Beatrix which give high ratings to science fiction movies (green shading in Figure 4.9). This suggests that $\boldsymbol{v_1}$ may reflect an idealized notion of a science fiction lover.

Similarly, $\boldsymbol{u}_2$, seems to capture a French art house film theme, and $\boldsymbol{v}_2$ may be reflecting that Chandra is close to an idealized lover of such movies. An idealized science fiction lover is a purist and only loves science fiction movies, so a science fiction lover $\boldsymbol{v}_1$ gives a rating of zero to everything but science fiction themed – this logic is implied by us requiring a diagonal substructure for the singular value matrix. A specific movie is therefore represented by how it decomposes (linearly) into its stereotypical movies. Likewise a person would be represented by how they decompose (via linear combination) into movie themes.

Remark It is worth discussing briefly SVD terminology and conventions as there are different versions used in the literature—the mathematics remains invariant to these differences— but can confuse the unaware reader:

- For convenience in notation and abstraction we use here an SVD notation where the SVD is described as having two square left- and right-

singular vector matrices, but a non-square singular value matrix. Our definition (4.74) for the SVD is sometimes called the *full SVD*. full SVD

- Some authors define the SVD a bit differently, for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $m \geqslant n$

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times n}{\boldsymbol{U}}\,\underset{n\times n}{\boldsymbol{\Sigma}}\,\underset{n\times n}{\boldsymbol{V}^T} \tag{4.107}$$

  Some authors call this the *reduced SVD* (e.g. Datta (2010)) other refer to this as *the* SVD (e.g. Press et al. (2007)). This alternative format changes merely how the matrices are constructed but leaves the mathematical structure of the SVD unchanged. The convenience of this alternative notation is that $\boldsymbol{\Sigma}$ is diagonal, as in the eigenvalue decomposition. However, it looses the interpretation of $\boldsymbol{\Sigma}$ as a transformation matrix. reduced SVD
- In Section 4.6, we will learn about matrix approximation techniques using the SVD, which is also called the *truncated SVD*. truncated SVD
- One can also define the SVD of a rank-r matrix $\boldsymbol{A}$ so that $\boldsymbol{U}$ is an $m \times r$ matrix, $\boldsymbol{\Sigma}$ as a diagonal matrix $r \times r$, and $\boldsymbol{V}$ as $r \times n$ matrix. This construction is very similar to our definition, and ensures that the diagonal matrix $\boldsymbol{\Sigma}$ has only non-zero entries along the diagonal. The main convenience of this alternative notation is that $\boldsymbol{\Sigma}$ is diagonal, as in the eigenvalue decomposition.
- One could also introduce the restriction that the SVD for $\boldsymbol{A}$ only applies to $m \times n$ matrices with $m > n$. However, this restriction is practically unnecessary. When $m < n$ the SVD decomposition will yield $\boldsymbol{\Sigma}$ with more zero columns than rows and, consequently, the singular values $\sigma_{m+1}, \ldots, \sigma_n$ are implicitly 0.

♢

The SVD is used in a variety of applications in machine learning from least squares problems in curve fitting to solving systems of linear equations. These applications harness various important properties of the SVD, its relation to the rank of a matrix and its ability to approximate matrices of a given rank with lower rank matrices. Substituting a matrix with the SVD form of that matrix in computations has often the advantage of making the calculation more robust to numerical rounding errors. As we will explore in the next section the SVD's ability to approximate matrices with "simpler" matrices in a principled manner opens up machine learning applications ranging from dimensionality reduction, topic modeling to data compression and clustering.

4.6 Matrix Approximation

We will now investigate how the SVD allows us to represent a matrix $\boldsymbol{A}$ as a sum of simpler matrices $\boldsymbol{A}_i$.

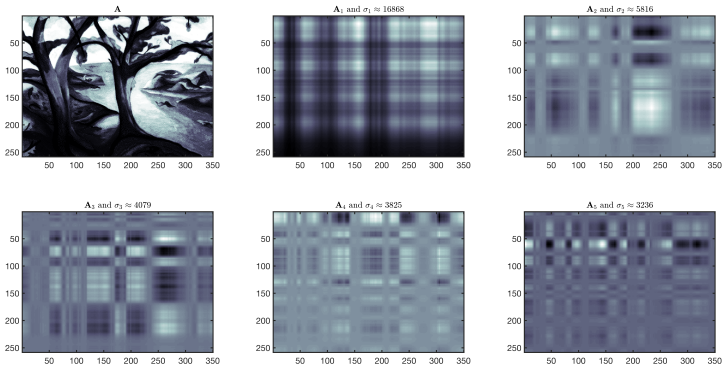


Figure 4.10 (Top left) A grayscale image is a 280×350 matrix of values between 0 (black) and 1 (white). (Middle Top to Bottom right) rank-1 matrices $\boldsymbol{A}_1 \dots \boldsymbol{A}_5$ and their corresponding singular values $\sigma_1, \dots, \sigma_5$. Note, that the grid like structure of each rank-1 matrix is imposed by the outer-product of the left and right singular vectors.

Let us construct a rank-1 $m \times n$ matrix $\boldsymbol{A}_i$ as

$$\boldsymbol{A}_i = \boldsymbol{u}_i \boldsymbol{v}_i^\top \tag{4.108}$$

which is formed by the outer product of ith orthogonal column vector of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively (see Figure 4.10 for a a visual example). For a matrix $\boldsymbol{A}$ of rank r the matrix can be decomposed into a sum of rank-1 matrices $\boldsymbol{A}_i$ as follows :

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i \tag{4.109}$$

where the outer product matrices $\boldsymbol{A}_i$ are weighed by the size of the ith singular value σ_i. Thus, the sum of the outer products of matching left and right singular vectors (weighted by their singular value) is equal to $\boldsymbol{A}$. Note, that any terms $i > r$ are zero, as the singular values will be 0. We can see why (4.108) holds: the diagonal structure of the singular value matrix $\boldsymbol{\Sigma}$ multiplies only matching left- and right-singular vectors $(\boldsymbol{u}_i, \boldsymbol{v}_i^\top)$ and adds them up, while setting non-matching left- and right-singular vectors $(\boldsymbol{u}_i, \boldsymbol{v}_j^\top, i \neq j)$ to zero.

In the previous paragraph we introduced a low-rank matrix $\boldsymbol{A}_i$ (of rank 1). We summed up the r individual rank-1 matrices to obtain a rank r matrix $\boldsymbol{A}$. What happens if the sum does not run over all matrices $\boldsymbol{A}_i$ from $i = 1 \dots r$ but instead run the sum only up to an intermediate value $k < r$. We are obtaining now an approximation of $\boldsymbol{A}$ that we call the *rank-k approximation* $\widehat{\boldsymbol{A}}(k)$

rank-k approximation

$$\widehat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \tag{4.110}$$

of $\boldsymbol{A}$ with $\mathrm{rk}(\widehat{\boldsymbol{A}}) = k$.

It would be useful if we could measure how large the difference between $\boldsymbol{A}$ and its approximation $\widehat{\boldsymbol{A}}(k)$ is in terms of a single number – we

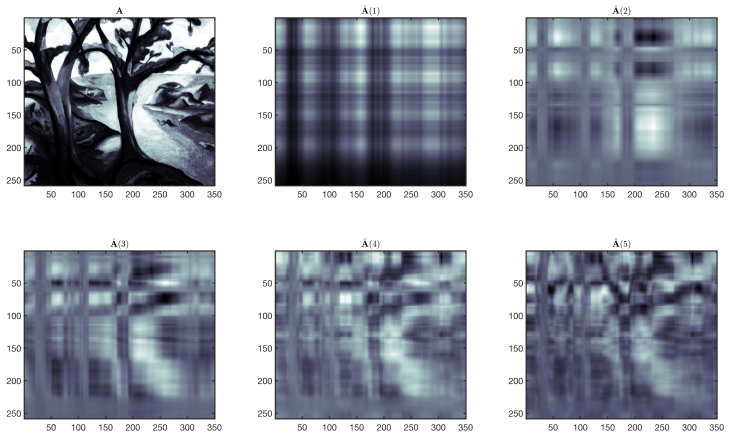


Figure 4.11 (Top left) The same grayscale image as in Figure 4.10. (Top Middle to Bottom right) Image reconstruction using the low-rank approximation of the SVD: (Top middle) is $\widehat{\boldsymbol{A}(1)} = \sigma_1 \boldsymbol{A}_1$. (Top right) is the rank-2 approximation $\widehat{\boldsymbol{A}(2)} = \sigma_1 \boldsymbol{A}_1 + \sigma_2 \boldsymbol{A}_2$. (Bottom left to Bottom right) are $\widehat{\boldsymbol{A}(3)}$ to $\widehat{\boldsymbol{A}(5)}$. Note how the shape of the trees becomes increasingly visible and clearly recognizable in the rank-6 approximation. While the original image requires $280 \times 350 = 98000$ numbers, the rank-6 approximation requires us only to store the 6 singular values and the 6 left and right singular vectors (280 and 350 dimensional each) for a total of $6 \times (280 + 350 + 1) = 3786$ numbers – just under 4% of the original.

thus need the notion of a norm. We have already used norms on vectors that measure the length of a vector. By analogy we can also define a norm on matrices (one of the many ways to define matrix norms).

Definition 4.23 (Spectral norm of a matrix) The spectral norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is defined as the following for $\boldsymbol{x} \in \mathbb{R}^n$

$$\|\boldsymbol{A}\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \quad \boldsymbol{x} \neq \boldsymbol{0}. \tag{4.111}$$

Note, we introduce the notation of a subscript in the matrix norm (left-hand side), by drawing on the notation for the euclidean norm for vectors (right-hand side) which has index 2.

The matrix spectral norm implies how long any vector $\boldsymbol{x}$ can at most become once it is multiplied by $\boldsymbol{A}$. This maximum lengthening is given by the SVD of $\boldsymbol{A}$.

Theorem 4.24 *The spectral norm of $\boldsymbol{A}$ is its largest singular value σ_1.*

We provide here a derivation of the largest singular value of matrix $\boldsymbol{A}$, illustrating the relation between the spectral norm and SVD.

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\boldsymbol{x}} \sqrt{\frac{\|\boldsymbol{A}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}} \tag{4.112}$$

$$= \max_{\boldsymbol{x}} \sqrt{\frac{(\boldsymbol{A}\boldsymbol{x})^\top (\boldsymbol{A}\boldsymbol{x})}{\boldsymbol{x}^\top \boldsymbol{x}}} = \max_{\boldsymbol{x}} \sqrt{\frac{\boldsymbol{x}^\top (\boldsymbol{A}^\top \boldsymbol{A}) \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}} \tag{4.113}$$

the matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is symmetric by construction and therefore we can compute the eigenvalue decomposition $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top$

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}} \sqrt{\frac{\boldsymbol{x}^\top(\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top)\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}}}\,, \tag{4.114}$$

$$\tag{4.115}$$

where $\boldsymbol{D}$ is a diagonal matrix containing the eigenvalues. Recall that $\boldsymbol{P}^\top$ and $\boldsymbol{P}$ perform merely a basis change and then undo it. Therefore, the most a vector $\boldsymbol{x}$ can be lengthened is if it is collinear with the eigenvector associated with the largest eigenvalue.

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_1} \tag{4.116}$$

the largest eigenvalue of $\boldsymbol{A}^\top\boldsymbol{A}$ is by (4.88) the largest singular value of $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_2 = \sigma_1 \tag{4.117}$$

Theorem 4.25 (Eckart-Young (or Eckart-Young-Minsky) theorem) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix of rank r and and $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ be a matrix of rank k. For any $k \leqslant r$ such that $\widehat{\boldsymbol{A}}(k) = \sum_i^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, it holds that*

$$\left\|\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)\right\|_2 = \sigma_{k+1} \tag{4.118}$$

$$= \min_{\mathrm{rk}(\boldsymbol{B})\leqslant k} \|\boldsymbol{A} - \boldsymbol{B}\|_2\,. \tag{4.119}$$

Remark We can interpret the rank-k approximation obtained with the SVD as a projection of the full rank matrix $\boldsymbol{A}$ onto the lower-dimensional space of rank at-most-k matrices. Of all possible projections the SVD rank-k approximation minimizes the difference with respect to the spectral norm between $\boldsymbol{A}$ and any rank-k matrix. ♢

We can retrace some of the steps to understand why (4.118) should hold. We observe that the difference between $\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)$ is a matrix containing the sum of the remaining rank-1 matrices

$$\boldsymbol{A} - \widehat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \tag{4.120}$$

Thus, by applying the definition of the spectral norm, (4.111), the largest amount a vector can be lengthened by the difference matrix is determined its largest singular value i.e. σ_{k+1}, which is the difference matrix's spectral norm.

Let us proceed to better understand (4.119) validity. We assume that there is another matrix $\boldsymbol{B}$ with $\mathrm{rk}(\boldsymbol{B}) \leqslant k$ such that

$$\|\boldsymbol{A} - \boldsymbol{B}\|_2 < \left\|\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)\right\|_2 \tag{4.121}$$

Then there exists an $(n-k)$-dimensional nullspace $Z \subseteq \mathbb{R}^n$ such that

$\boldsymbol{x} \in Z \implies \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$. In other words, we have an n-dimensional space $\mathbb{R}^n$ in which lies a lower dimensional nullspace of $\boldsymbol{B}$. Then it follows that

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2 \ , \tag{4.122}$$

and by using a version of the Cauchy-Schwartz inequality (3.17) that encompasses norms of matrices we obtain

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\boldsymbol{x}\|_2 < \sigma_{k+1} \|\boldsymbol{x}\|_2 \tag{4.123}$$

We obtain an $(n-k)$ dimensional subspace where $\|\boldsymbol{A}\boldsymbol{x}\|_2 < \sigma_{k+1} \|\boldsymbol{x}\|_2$.

On the other hand there exists a $(k+1)$-dimensional subspace where $\|\boldsymbol{A}\boldsymbol{x}\|_2 \geqslant \sigma_{k+1} \|\boldsymbol{x}\|_2$ which is spanned by the right singular vectors $\boldsymbol{v}_j, j \leqslant k+1$ of $\boldsymbol{A}$. Adding up dimensions of these two spaces yields a number greater n, as there must be a non-zero vector in both spaces. This is a contradiction because of the Rank-Nullity Theorem (recall Theorem 2.23 in Section 2.7.3).

The Eckart-Young theorem implies that we can use SVD to reduce a rank-r matrix $\boldsymbol{A}$ to a rank-k matrix $\widehat{\boldsymbol{A}}$ in a principled, optimal (in the spectral norm sense) manner. The effect of the low-rank approximation is that we can obtain a more compact representation of the values of the matrix with limited loss of information, this is a form of data compression. Therefore, the low-rank approximation of a matrix appears in many machine learning applications, such as image processing, noise filtering, and regularization of ill-posed problems. Furthermore, it plays a key role in dimensionality reduction and principal component analysis as we shall see in Chapter 10.

Example 4.14 (Finding Structure in Movie Ratings and Consumers (continued))

Following from our previous movie rating example we can now apply the concept of low-rank approximation to describe the data matrix. Recall that our first singular value captures the notion of science fiction theme in movies and science fiction lovers. Thus, by using only the first singular value term in a rank-1 decomposition of the movie rating matrix we obtain the following predicted ratings

$$\boldsymbol{M}_1 = \sigma_1(\boldsymbol{u}_1\boldsymbol{v}_1^\top) \tag{4.124}$$

$$= 9.6438 \begin{bmatrix} -0.6710 \\ -0.7197 \\ -0.0939 \\ -0.1515 \end{bmatrix} \begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \end{bmatrix} \tag{4.125}$$

$$= \begin{bmatrix} 4.7673 & 4.2154 & 1.1718 \\ 5.1138 & 4.5218 & 1.2570 \\ 0.6671 & 0.5899 & 0.1640 \\ 1.0765 & 0.9519 & 0.2646 \end{bmatrix} \tag{4.126}$$

This first rank-1 approximation $\boldsymbol{M}_1$ is insightful: it tells us that Ali and Beatrix like science fiction movies such as Star Wars and Bladerunner (entries have values > 4), but on the other hand fails to capture the ratings of the other movies by Chandra. This is not surprising as Chandra's type of movies are not captured by the first singular value. The second singular value however gives us a better rank-1 approximation for those movie theme-movie lovers types.

$$\boldsymbol{M}_2 = \sigma_2(\boldsymbol{u}_2\boldsymbol{v}_2^\top) \tag{4.127}$$

$$= 6.3639 \begin{bmatrix} 0.0236 \\ 0.2054 \\ -0.7705 \\ -0.6030 \end{bmatrix} \begin{bmatrix} 0.0852 & 0.1762 & -0.9807 \end{bmatrix} \tag{4.128}$$

$$= \begin{bmatrix} 0.0128 & 0.0265 & -0.1475 \\ 0.1114 & 0.2304 & -1.2820 \\ -0.4178 & -0.8642 & 4.8084 \\ -0.3270 & -0.6763 & 3.7631 \end{bmatrix} \tag{4.129}$$

In this second rank-1 approximation $\boldsymbol{M}_2$ we capture Chandra's ratings and movie types well, but for the science fiction movies and people the predictions are, not surprisingly, poor.

This leads us to consider the rank-2 approximation $\widehat{\boldsymbol{A}}(2)$ where we combine the first two rank-1 approximations

$$\widehat{\boldsymbol{A}}(2) = \boldsymbol{M}_1 + \boldsymbol{M}_2 \tag{4.130}$$

$$= \begin{bmatrix} 4.7801 & 4.2419 & 1.0244 \\ 5.2252 & 4.7522 & -0.0250 \\ 0.2493 & -0.2743 & 4.9724 \\ 0.7495 & 0.2756 & 4.0278 \end{bmatrix} \tag{4.131}$$

$\widehat{\boldsymbol{A}}(2)$ is close to the original movie ratings table

$$\boldsymbol{A} = \begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} \tag{4.132}$$

and this suggests that we can ignore the third singular value (after all it is much smaller than the first two). We can interpret this as to imply that in the data table there really is no evidence of a third movie-theme-movie lovers category. This also means that the entire space of movie themes-movie lovers is in our example a two-dimensional space spanned by science fiction and French art house movies and lovers.

4.7 Matrix Phylogeny

In Chapter 2 and 3 we covered the basics of linear algebra and analytic geometry, in this chapter we now looked at fundamental characteristics and methods on matrices and linear mappings. We are depicting in Figure 4.12 the phylogenetic tree of relationships between different types of matrices (black arrows indicating "is a subset of") and the covered operations we can perform on them (in red). For example, we already learned in Chapter 2 about **square** matrices, which are a subset of **all (complex) matrices** (top level node in the tree). We will then learn here that we can compute a specific characteristic (**determinant**) in Section 4.1 that will inform us whether a square matrix has an associate **inverse matrix**, thus if it belongs to the class of non-singular, invertible matrices.

The word phylogenetic describes how we capture the relationships among individuals or groups and derived from the greek words for "tribe" and "source".

Going backward through the chapter, we start with the most general case of real matrices $\mathbb{R}^{n\times m}$ for which we can define a pseudo-inverse to "invert" them, as well as perform **singular value decomposition (SVD)** (Theorem 4.22). This superset of matrices is divided into the square $\mathbb{R}^{n\times n}$ matrices for which we can define the characteristic feature of the **determinant** and the **trace** (Section 4.1).

Here the set of matrices splits in two: If the square $\mathbb{R}^{n\times n}$ matrix has n distinct eigenvalues (or equivalently n linearly independent eigenvectors) then the matrix is non-defective and a unique **diagonalisation/eigendecomposition** exists for these matrices (Theorem 4.13). In other cases we know that a multiplicity of eigenvalues may result (see Definitions 4.10 and 4.11).

Alternatively, if this square $\mathbb{R}^{n\times n}$ matrix has a non-zero determinant, than the matrix is non-singular, i.e. an inverse matrix exists (Theorem 4.1). Non-singular matrices are closed under addition and multiplication, have an identity element ($\boldsymbol{I}$) and an inverse element, thus they form a group.

Note, that non-singular and non-defective matrices are not identical sets, as for example a rotation matrix will be invertible (determinant is non-zero) but not diagonalizable in the real numbers (eigenvalues are not real numbers).

We dive further into the branch of non-defective square $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ matrices. $\boldsymbol{A}$ is normal if the condition $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top$ holds. Moreover, if the more restrictive condition holds $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}$, then the matrix is called orthogonal (see Definition 3.8) and is a subset of the non-singular (invertible) matrices and satisfy the very useful condition $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$. Orthogonal matrices are closed under addition and multiplication, have an identity element ($\boldsymbol{I}$) and an inverse element, thus they also form a group.

The normal matrices have a frequently encountered subset, the symmetric matrices $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ which satisfy $\boldsymbol{S} = \boldsymbol{S}^\top$. Symmetric matrices have only real eigenvalues. A subset of the symmetric matrices are the positive definite matrices $\boldsymbol{P}$ that satisfy the condition of $\boldsymbol{x}^\top\boldsymbol{P}\boldsymbol{x} > 0$, then

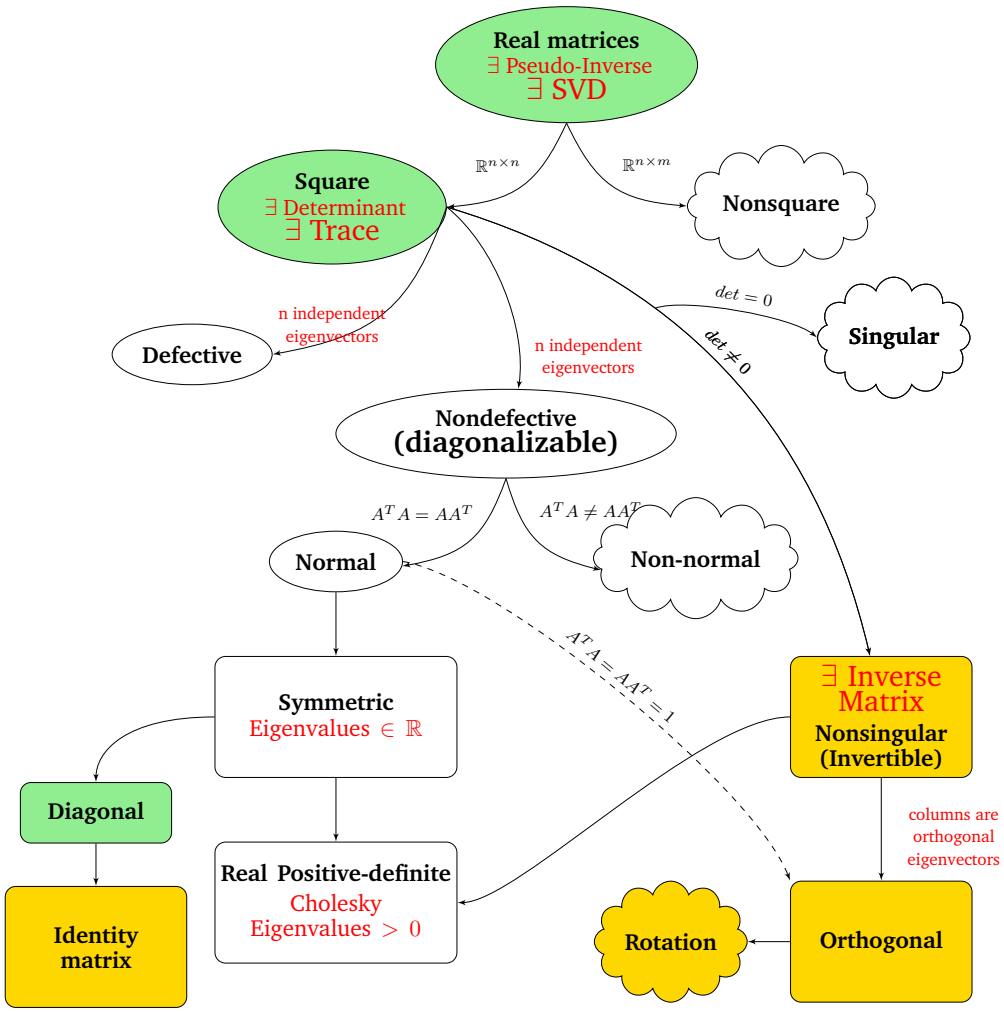


Figure 4.12 A functional phylogeny of matrices encountered in machine learning.

a unique **Cholesky decomposition** exists (Theorem 4.18). Positive definite matrices have only positive eigenvalues and are always invertible (i.e. have a non-zero determinant).

Another subset of the symmetric matrices are the **diagonal matrices** $\boldsymbol{D}$ in which the entries outside the main diagonal are all zero. Diagonal matrices are closed under multiplication and addition, but do not necessarily form a group (this is only the case if all diagonal entries are non-zero so that the matrix is invertible). A prominent special case of the diagonal matrices is the identity matrix $\boldsymbol{I}$.

4.8 Further Reading

Most of the content in this chapter establishes underlying mathematics and connects them to methods for studying mappings, many of these underly machine learning at the level of underpinning software solutions and building blocks for almost all machine learning theory. Matrix characterization using determinants, eigenspectra and eigenspaces are fundamental features and conditions for categorizing and analyzing matrices, this extends to all forms of representations of data and mappings involving data, as well as judging the numerical stability of computational operations on such matrices (Press et al., 2007).

Determinants are fundamental tools in order to invert matrices and compute eigenvalues "by hand", yet for almost all but the smallest instances numerical computation by Gaussian elimination outperforms determinants (Press et al., 2007). Determinants remain however a powerful theoretical concept, e.g. to gain intuition about the orientation of a basis based on the sign of the determinant. Eigenvectors can be used to perform change of basis operations so as to transform complicated looking data into more meaningful orthogonal, feature vectors. Similarly, matrix decomposition methods such as Cholesky decomposition reappear often when we have to compute or simulate random events (Rubinstein and Kroese, 2016). Therefore, the Cholesky decomposition enables us to compute the *reparametrization trick* where we want to perform continuous differentiation over random variables, e.g., in variational autoencoders Kingma and Welling (2013); Jimenez Rezende et al. (2014).

Eigendecomposition is fundamental in enabling us to extract meaningful and interpretable information that characterizes linear mappings. Therefore, eigendecomposition underlies a general class of machine learning algorithms called *spectral methods* that perform eigendecomposition of a positive-definite kernel. These spectral decomposition methods encompass classical approaches to statistical data analysis, such as

- Principal Components Analysis (PCA (Pearson, 1901), see also Chapter 10), in which a low-dimensional subspace that explains most of the variability in the data is sought.

- Fisher Discriminant Analysis, which aims to determine a separating hyperplane for data classification (Mika et al., 1999).
- Multidimensional Scaling (MDS) (Carroll and Chang, 1970).

The computational efficiency of these methods typically results from finding the best rank-k approximation to a symmetric, positive semidefinite matrix. More contemporary examples of spectral methods have different origins, but each of them requires the computation of the eigenvectors and eigenvalues of a positive-definite kernel, such as

- Isomap (Tenenbaum et al., 2000),
- Laplacian eigenmaps (Belkin and Niyogi, 2003),
- Hessian eigenmaps (Donoho and Grimes, 2003),
- Spectral clustering (Shi and Malik, 2000).

The core computations of these are generally underpinned by low-rank matrix approximation techniques (Belabbas and Wolfe, 2009), as we encountered here via the SVD.

The SVD allows us to discover some of the same kind of information as the eigendecomposition. However, the SVD is more generally applicable to non-square matrices, such as tables of data. These matrix factorisation methods become relevant whenever we want to identify heterogeneity in data when we want to perform data compression by approximation, e.g. instead of storing $n \times m$ values just storing $(n+m) \times k$ values, or when we want to perform data preprocessing, e.g. to decorrelate predictor variables of a design matrix (e.g. Ormoneit et al. (2001)). SVD operates on matrices and matrices are rectangular arrays that have two indices (rows and columns). The extension of matrix-like structure to higher-dimensional arrays are called tensors. It turns out that the SVD is the special case of a more general family of decompositions that operate on such tensors (Kolda and Bader, 2009). SVD-like operations and low-rank approximations on tensors are for example the Tucker Decomposition (Tucker, 1966) or the CP decomposition (Carroll and Chang, 1970).

The SVD low-rank approximation is frequently used in machine learning for computational efficiency reasons. This is because it reduces the amount of memory and operations with non-zero multiplications we need to perform on potentially very large matrices of data (Trefethen and Bau III, 1997). Moreover, low-rank approximation is used to operate on matrices that may contain missing values as well as for purposes of lossy compression and dimensionality reduction (Moonen and De Moor, 1995; Markovsky, 2011).

Exercises

4.1 Compute the determinant using the Laplace expansion (using the first row) and the Sarrus Rule for

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \\ 0 & 2 & 4 \end{bmatrix} \tag{4.133}$$

4.2 Compute the following determinant efficiently.

$$\begin{bmatrix} 2 & 0 & 1 & 2 & 0 \\ 2 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -2 & 0 & 2 & -1 & 2 \\ 2 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.134}$$

4.3 Compute the eigenspaces of $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $\begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$

4.4 Compute the eigenspaces of

$$\boldsymbol{A} = \begin{bmatrix} 0 & -1 & 1 & 1 \\ -1 & 1 & -2 & 3 \\ 2 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 \end{bmatrix} \tag{4.135}$$

4.5 Diagonalizability of a matrix is unrelated to its invertibility. Determine for the following four matrices it if is diagonalizable and/or invertible $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

4.6 Find the SVD of the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \tag{4.136}$$

4.7 Let us find the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} . \tag{4.137}$$

4.8 Find the rank-1 approximation of

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \tag{4.138}$$