

4

1911

Matrix Decompositions

1912 In Chapters 2 and 3, we studied ways to manipulate and measure vectors,
1913 projections of vectors and linear mappings. Mappings and transformations
1914 of vectors can be conveniently described as operations performed on ma-
1915 trices. Moreover, data is often represented in matrix form as well, for ex-
1916 ample where the rows of the matrix represent different instances of the
1917 data (for example people) and the columns describe different features of
1918 the data (for example weight, height and socio-economic status). In this
1919 chapter we present three aspects of matrices: how to summarize matrices,
1920 how matrices can be decomposed, and how these decompositions can be
1921 used to consider matrix approximations.

1922 We first consider methods that allow us to describe matrices with just
1923 a few numbers that characterize the overall properties of matrices. We
1924 will do this in the sections on determinants (Section 4.1) and eigenvalues
1925 (Section 4.2 for the important special case of square matrices. These char-
1926 acteristic numbers have important mathematical consequences and allow
1927 us to quickly grasp what useful properties a matrix has. From here we will
1928 proceed to matrix decomposition methods: An analogy for matrix decom-
1929 position is the factoring of numbers, such as the factoring of 21 into prime
1930 numbers 7×3 . For this reason matrix decomposition is also often referred
matrix factorization 1931 to as *matrix factorization*. Matrix decompositions are used to interpret a
1932 matrix using a different representation using factors of interpretable ma-
1933 trices.

1934 We will first cover a square-root-like operation for matrices called Cholesky
1935 decomposition (Section 4.3) for symmetric, positive definite matrices. From
1936 here we will look at two related methods for factorizing matrices into
1937 canonical forms. The first one is known as matrix diagonalization (Sec-
1938 tion 4.4), which allows us to represent the linear mapping using a diag-
1939 onal transformation matrix if we choose an appropriate basis. The second
1940 method, singular value decomposition (Section 4.5), extends this factor-
1941 ization to non-square matrices, and it is considered one of the fundamen-
1942 tal concepts in linear algebra. These decomposition are helpful as matrices
1943 representing numerical data are often very large and hard to analyze. We
1944 conclude the chapter with a systematic overview of the types of matrices
1945 and the characteristic properties that distinguish them in form of a matrix
1946 taxonomy (Section 4.7).

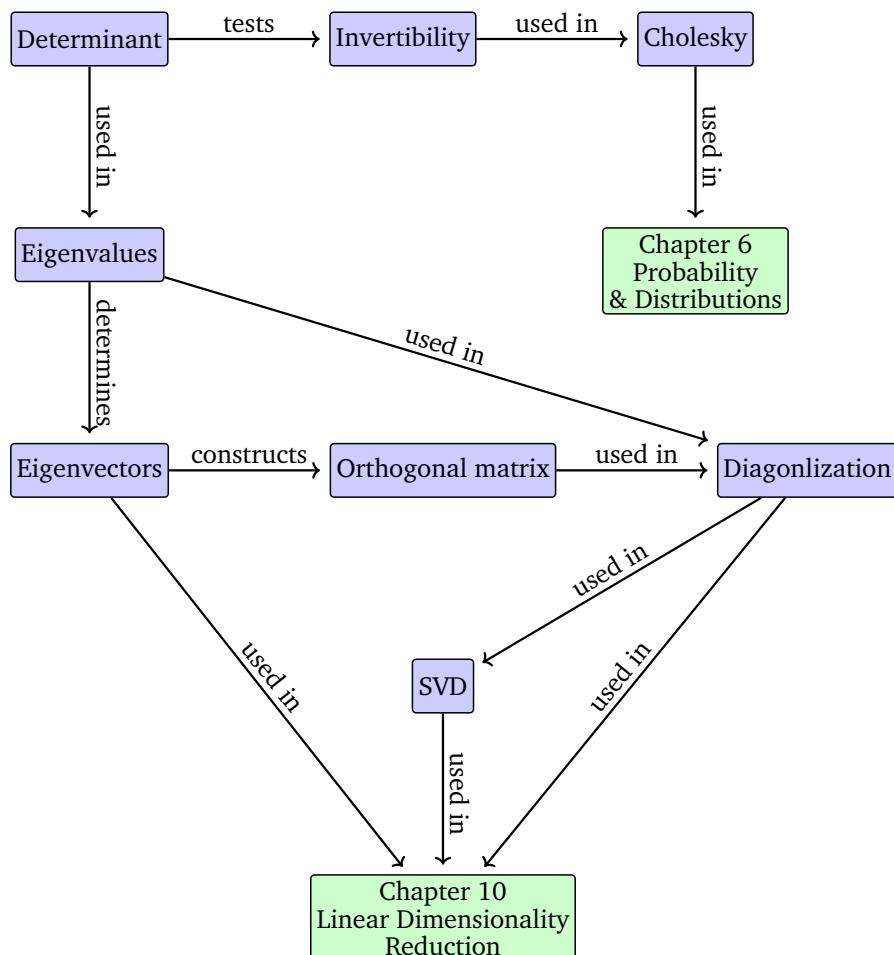


Figure 4.1 A mind map of the concepts introduced in this chapter, along with when they are used in other parts of the book.

1947 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.

1952

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 $\mathbf{A} \in \mathbb{R}^{n \times n}$, that is matrices with the same number of rows and columns. In this book we write this as $\det(\mathbf{A})$ (some textbooks may use $|\mathbf{A}|$, 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} be

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

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

determinant 1953 The determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a function that maps
 1954 \mathbf{A} onto a real number. Before provide a definition of the determinant for
 1955 general $n \times n$ matrices let us look at some motivating examples, and define
 1956 determinants for some special matrices.

Example 4.1 (Testing for Matrix Invertibility)

Let us begin with exploring if a square matrix \mathbf{A} is invertible (see Section 2.2.2). For the smallest cases, we already know when a matrix is invertible. If \mathbf{A} is a 1×1 matrix, i.e., it is a scalar number, then $\mathbf{A} = a \implies \mathbf{A}^{-1} = \frac{1}{a}$. 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 $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ and thus we can write that the inverse of \mathbf{A}^{-1} is (from Equation 2.23)

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \quad (4.2)$$

Thus, \mathbf{A} is invertible if and only if

$$a_{11}a_{22} - a_{12}a_{21} \neq 0. \quad (4.3)$$

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

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

1957 The example above points already at the relationship between determinants
 1958 and the existence of inverse matrices. The next theorem states the
 1959 same result for $n \times n$ matrices.

1960 **Theorem 4.1.** For any square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ it holds that \mathbf{A} is invertible
 1961 if and only if $\det(\mathbf{A}) \neq 0$.

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

$$\det(\mathbf{A}) = \det(a_{11}) = a_{11}. \quad (4.5)$$

For $n = 2$,

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

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

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

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 \mathbf{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 \mathbf{A} , the determinant is the product of the diagonal elements:

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

upper triangular matrix
lower triangular matrix

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 \mathbf{A} .

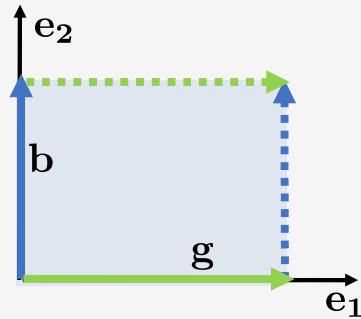


Figure 4.2
Determinants can measure areas spanned by vectors. The area A of the parallelogram (shaded region) spanned by the vectors \mathbf{b} and \mathbf{g} is given by the determinant $\det([\mathbf{b}, \mathbf{g}])$.

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 \mathbf{b}, \mathbf{g} that form the columns of a matrix $\mathbf{A} = [\mathbf{b}, \mathbf{g}]$. Then, the absolute value of the determinant of \mathbf{A} is the area of the parallelogram with vertices $0, \mathbf{b}, \mathbf{g}, \mathbf{b} + \mathbf{g}$. In particular, if the two vectors \mathbf{b}, \mathbf{g} were linearly dependent so that $\mathbf{b} = \lambda\mathbf{g}$ for some $\lambda \in \mathbb{R}$ they no longer form a two-dimensional parallelogram. Therefore, the corresponding area is 0. On the contrary, if \mathbf{b}, \mathbf{g} were al and lie along the canonical coordinate axes e_1, e_2 then they

would reduce to $\mathbf{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$ and $\mathbf{g} = \begin{bmatrix} 0 \\ g \end{bmatrix}$ and the determinant

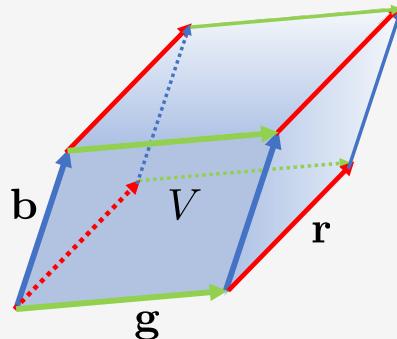
$$\begin{vmatrix} b & 0 \\ 0 & g \end{vmatrix} = bg - 0 = bg \quad (4.9)$$

becomes the familiar formula: area = height \times length.

The sign of the determinant measures the orientation of the spanning vectors \mathbf{b}, \mathbf{g} with respect to the standard coordinate system e_1, e_2 . In our figure, flipping the spanning order to \mathbf{g}, \mathbf{b} swaps the columns of A and reverses the orientation of the shaded surface A .

This intuition extends to higher dimensions. In \mathbb{R}^3 , we consider three vectors $\mathbf{r}, \mathbf{b}, \mathbf{g} \in \mathbb{R}^3$ 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 $[\mathbf{r}, \mathbf{b}, \mathbf{g}]$ 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 $\mathbf{r}, \mathbf{b}, \mathbf{g}$ is given by the determinant $\det([\mathbf{r}, \mathbf{b}, \mathbf{g}])$.



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

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

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

Therefore, the volume is given as

$$V = |\det(\mathbf{A})| = 186. \quad (4.12)$$

¹⁹⁶²

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 deter-

¹⁹⁶³

¹⁹⁶⁴

1965 minant of an $n \times n$ matrix to computing the determinant of $(n-1) \times (n-1)$
 1966 matrices. By recursively applying the Laplace expansion we can therefore
 1967 compute determinants of $n \times n$ matrices by ultimately computing deter-
 1968 minants of 2×2 matrices.

1969 **Theorem 4.2** (Laplace Expansion). *Consider a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then,*
 1970 *for all $j = 1, \dots, n$:*

1. *Expansion along column j*

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

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

2. *Expansion along row j*

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

1971 Here $\mathbf{A}_{k,j} \in \mathbb{R}^{(n-1) \times (n-1)}$ is the submatrix of \mathbf{A} that we obtain when delet-
 1972 ing row k and column j .

Example 4.3 (Laplace Expansion)

Let us compute the determinant of

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \quad (4.15)$$

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

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

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

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

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

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

1973 For $\mathbf{A} \in \mathbb{R}^{n \times n}$ the determinant exhibits the following properties:

- 1974 • The determinant of a product is the product of the determinant, $\det(\mathbf{AB}) =$
 1975 $\det(\mathbf{A})\det(\mathbf{B})$.

- 1976 • Determinants are invariant to transposition $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$.
- 1977 • If \mathbf{A} is regular (Section 2.2.2) then $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}$
- 1978 • Similar matrices (Defintion 2.21) possess the same determinant. Therefore, for a linear mapping $\Phi : V \rightarrow V$ all transformation matrices \mathbf{A}_Φ of Φ have the same determinant. Thus, the determinant is invariant to the choice of basis of a linear mapping.
- 1980 • Adding a multiple of a column/row to another one does not change $\det(\mathbf{A})$.
- 1982 • Multiplication of a column/row with $\lambda \in \mathbb{R}$ scales $\det(\mathbf{A})$ by λ . In particular, $\det(\lambda\mathbf{A}) = \lambda^n \det(\mathbf{A})$.
- 1984 • Swapping two rows/columns changes the sign of $\det(\mathbf{A})$.

1987 Because of the last three properties, we can use Gaussian elimination (see
 1988 Section 2.1) to compute $\det(\mathbf{A})$ by bringing \mathbf{A} into row-echelon form. We
 1989 can stop Gaussian elimination when we have \mathbf{A} in a triangular form where
 1990 the elements below the diagonal are all 0. Recall from Equation (4.8) that
 1991 the determinant is then the product of the diagonal elements.

1992 **Theorem 4.3.** A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has $\det(\mathbf{A}) \neq 0$ if and only if
 1993 $\text{rk } \mathbf{A} = n$. In other words a square matrix is invertible if and only if it is full
 1994 rank.

1995 When mathematics was mainly performed by hand, the determinant
 1996 calculation was considered an essential way to analyze matrix invertibility.
 1997 However, contemporary approaches in machine learning use direct
 1998 numerical methods that superseded the explicit calculation of the deter-
 1999 minant. For example, in Chapter 2 we learned that inverse matrices can
 2000 be computed by Gaussian elimination. Gaussian elimination can thus be
 2001 used to compute the determinant of a matrix.

2002 Determinants will play an important theoretical role for the following
 2003 sections, especially when we learn about eigenvalues and eigenvectors
 2004 (Section 4.2) through the characteristic polynomial of a matrix.

trace

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

$$\text{tr}(\mathbf{A}) := \sum_{i=1}^n a_{ii}, \quad (4.18)$$

2005 in other words, the trace is the sum of the diagonal elements of \mathbf{A} .

2006 *Remark.* For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ the trace satisfies the following properties:

- 2007 1. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
- 2008 2. $\text{tr}(\alpha\mathbf{A}) = \alpha\text{tr}(\mathbf{A}), \quad \alpha \in \mathbb{R}$
- 2009 3. $\text{tr}(\mathbf{I}_n) = n$
- 2010 4. $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$

2011 It can be shown that only one function satisfies these four properties together – the trace (Gohberg et al., 2012). \diamond

2013 *Remark.* The properties of the trace of matrix products are more general:

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

$$\text{tr}(\mathbf{A}\mathbf{K}\mathbf{L}) = \text{tr}(\mathbf{K}\mathbf{L}\mathbf{A}) \quad (4.19)$$

2014 for matrices $\mathbf{A} \in \mathbb{R}^{a \times k}$, $\mathbf{K} \in \mathbb{R}^{l \times l}$, $\mathbf{L} \in \mathbb{R}^{l \times a}$. 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 $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$:

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (4.20)$$

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

$$\text{tr}(\mathbf{xy}^\top) = \text{tr}(\mathbf{y}^\top \mathbf{x}) = \mathbf{y}^\top \mathbf{x} \in \mathbb{R}. \quad (4.21)$$

\diamond

Remark. Given some linear map $\Phi : V \rightarrow V$, 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 \mathbf{A} relative to this basis, and taking the trace of this square matrix. Assume that \mathbf{B} is transformation matrix between bases of V . Then, we can write

$$\text{tr}(\mathbf{BAB}^{-1}) = \text{tr}(\mathbf{B}^{-1}\mathbf{BA}) = \text{tr}(\mathbf{IA}) = \text{tr}(\mathbf{A}). \quad (4.22)$$

2017 Thus, while matrix representations of linear mappings are basis dependent 2018 its trace is independent of the basis. \diamond

2019 The trace is useful in certain classes of machine learning models where 2020 data is fitted using linear regression models. The trace captures model 2021 complexity in these models and can be used to compare between models 2022 (a more principled foundation for model comparison is discussed in detail 2023 in Section 8.5).

2024 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 \mathbf{A} in terms of a polynomial, which we will use extensively in the 2025 following sections.

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

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) \quad (4.23)$$

$$= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \quad (4.24)$$

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

$$c_0 = \det(\mathbf{A}), \quad (4.25)$$

$$c_{n-1} = (-1)^{n-1}\text{tr}(\mathbf{A}). \quad (4.26)$$

characteristic
polynomial

2029 The characteristic polynomial will allow us to compute eigenvalues and
2030 eigenvectors, covered in the next section.

2031 **4.2 Eigenvalues and Eigenvectors**

2032 We will now get to know a new way to characterize a matrix and, its as-
2033 sociated linear mapping. Let us recall from Section 2.7.1 that every linear
2034 mapping has a unique transformation matrix given an ordered basis. We
2035 can interpret linear mappings and their associated transformation matri-
2036 ces by performing an “Eigen” analysis. *Eigen* is a German word meaning
2037 “characteristic”, “self” or “own”. As we will see the eigenvalues of a lin-
2038 ear mapping will tell us how a special set of vectors, the eigenvectors, are
2039 transformed by the linear mapping.

2040 **Definition 4.6.** Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an
2041 eigenvalue of A and a nonzero $x \in \mathbb{R}^n$ is the corresponding eigenvector of
2042 A if

$$Ax = \lambda x. \quad (4.27)$$

eigenvalue
eigenvector

2043 eigenvalue equation We call this the *eigenvalue equation*.

2044 **Remark.** In linear algebra literature and software, it is often a conven-
2045 tion that eigenvalues are sorted in descending order, so that the largest
2046 eigenvalue and associated eigenvector are called the first eigenvalue and
2047 its associated eigenvector, and the second largest called the second eigen-
2048 value and its associated eigenvector, and so on. However textbooks and
2049 publications may have different or no notion of orderings. We do not want
2050 to presume an ordering in our book. ◇

codirected
collinear

2048 **Definition 4.7 (Collinearity & Codirection).** Two vectors that point in the
2049 same direction are called *codirected*. Two vectors are *collinear* if they point
2050 in the same or the opposite direction.

2051 **Remark (Non-uniqueness of Eigenvectors).** If x is an eigenvector of A
2052 associated with eigenvalue λ then for any $c \in \mathbb{R} \setminus \{0\}$ it holds that cx is
2053 an eigenvector of A with the same eigenvalue since

$$A(cx) = cAx = c\lambda x = \lambda(cx). \quad (4.28)$$

2054 Thus, all vectors that are collinear to x are also eigenvectors of A . ◇

eigenspace

2055 **Theorem 4.8.** $\lambda \in \mathbb{R}$ is eigenvalue of $A \in \mathbb{R}^{n \times n}$ if and only if λ is a root of
2056 the characteristic polynomial $p_A(\lambda)$ of A .

2057 **Definition 4.9 (Eigenspace and Eigenspectrum).** For $A \in \mathbb{R}^{n \times n}$ the set
2058 of all eigenvectors of A associated with an eigenvalue λ spans a subspace
2059 of \mathbb{R}^n , which is called *eigenspace* of A with respect to λ and is denoted

by E_λ . The set of all eigenvalues of \mathbf{A} is called the *eigenspectrum*, or just spectrum, of \mathbf{A} .

There are a number of ways to think about these characteristics

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

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \iff \mathbf{A}\mathbf{x} - \lambda\mathbf{x} = \mathbf{0} \quad (4.29)$$

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

- 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, the 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.

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}. \quad (4.31)$$

Step 1: Characteristic Polynomial

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

Step 2: Eigenvalues

The characteristic polynomial is given as

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \det \left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = \begin{vmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{vmatrix} \quad (4.32)$$

$$= (4 - \lambda)(3 - \lambda) - 2 \cdot 1. \quad (4.33)$$

We factorize the characteristic polynomial

$$p(\lambda) = (4 - \lambda)(3 - \lambda) - 2 \cdot 1 = 10 - 7\lambda + \lambda^2 = (2 - \lambda)(5 - \lambda) \quad (4.34)$$

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 \mathbf{x} such that

$$\begin{bmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{bmatrix} \mathbf{x} = \mathbf{0}. \quad (4.35)$$

For $\lambda = 5$ we obtain

$$\begin{bmatrix} 4 - 5 & 2 \\ 1 & 3 - 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}. \quad (4.36)$$

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

$$E_5 = \text{span} \left[\begin{bmatrix} 2 \\ 1 \end{bmatrix} \right], \quad (4.37)$$

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) but that in other cases where we have multiple eigenvalues (see Definition 4.13) the eigenspace may have more than one dimension.

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

$$\begin{bmatrix} 4 - 2 & 2 \\ 1 & 3 - 2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \mathbf{x} = \mathbf{0}. \quad (4.38)$$

This means any vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 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

$$E_2 = \text{span} \left[\begin{bmatrix} 1 \\ -1 \end{bmatrix} \right]. \quad (4.39)$$

2069 2070 2071 2072 2073 2074 2075 2076

Remark (Eigenvalues and Eigenspaces). If λ is an eigenvalue of $\mathbf{A} \in \mathbb{R}^{n \times n}$ then the corresponding eigenspace E_λ is the solution space of the homogeneous linear equation system $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$. 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 does not change its direction under \mathbf{A} . \diamond

2077 *Remark.* The following statements are equivalent:

- 2078 • λ is eigenvalue of $\mathbf{A} \in \mathbb{R}^{n \times n}$
- 2079 • There exists an $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ with $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ or equivalently, $(\mathbf{A} - \lambda \mathbf{I}_n)\mathbf{x} = \mathbf{0}$ can be solved non-trivially, i.e., $\mathbf{x} \neq \mathbf{0}$.

- 2081 • $\text{rk}(\mathbf{A} - \lambda \mathbf{I}_n) < n$
 2082 • $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$

◊

2083 Useful properties regarding eigenvalues and eigenvectors of various ma-
 2084 trix types include

- 2085 • A matrix \mathbf{A} and its transpose \mathbf{A}^\top possess the same eigenvalues, but not
 2086 necessarily the same eigenvectors.
 2087 • Symmetric matrices always have real-valued eigenvalues.
 2088 • Symmetric positive definite matrices always have positive, real eigen-
 2089 values.
 2090 • The eigenvectors of symmetric matrices are always orthogonal to each
 2091 other.

Theorem 4.10. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ we can always obtain a \mathbf{S} that is a symmetric positive semi-definite matrix by computing*

$$\mathbf{S} = \mathbf{A}^\top \mathbf{A}. \quad (4.40)$$

2093 Understanding why this theorem holds is insightful for how we can
 2094 use symmetrised matrices: Symmetry requires $\mathbf{S} = \mathbf{S}^\top$ and by inserting
 2095 (4.40) we obtain $\mathbf{S} = \mathbf{A}^\top \mathbf{A} = \mathbf{A}^\top (\mathbf{A}^\top)^\top = (\mathbf{A}^\top \mathbf{A})^\top = \mathbf{S}^\top$. More-
 2096 over, positive semi-definiteness (Section 3.2.3) requires that $\mathbf{x}^\top \mathbf{S} \mathbf{x} \geq 0$
 2097 and inserting (4.40) we obtain $\mathbf{x}^\top \mathbf{S} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} = (\mathbf{x}^\top \mathbf{A}^\top)(\mathbf{A} \mathbf{x}) =$
 2098 $(\mathbf{A} \mathbf{x})^\top (\mathbf{A} \mathbf{x}) \geq 0$, because the scalar product computes a sum of squares
 2099 (which are themselves always positive or zero).

2100 **Theorem 4.11** (Hogben (2006)). *Consider a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with distinct eigenvalues $\lambda_1, \dots, \lambda_n$ and corresponding eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$. Then the eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ are linearly independent.*

2103 The theorem states that eigenvectors belonging to different eigenvalues
 2104 form a linearly independent set. For symmetric matrices we can state a
 2105 stronger version of Theorem 4.11.

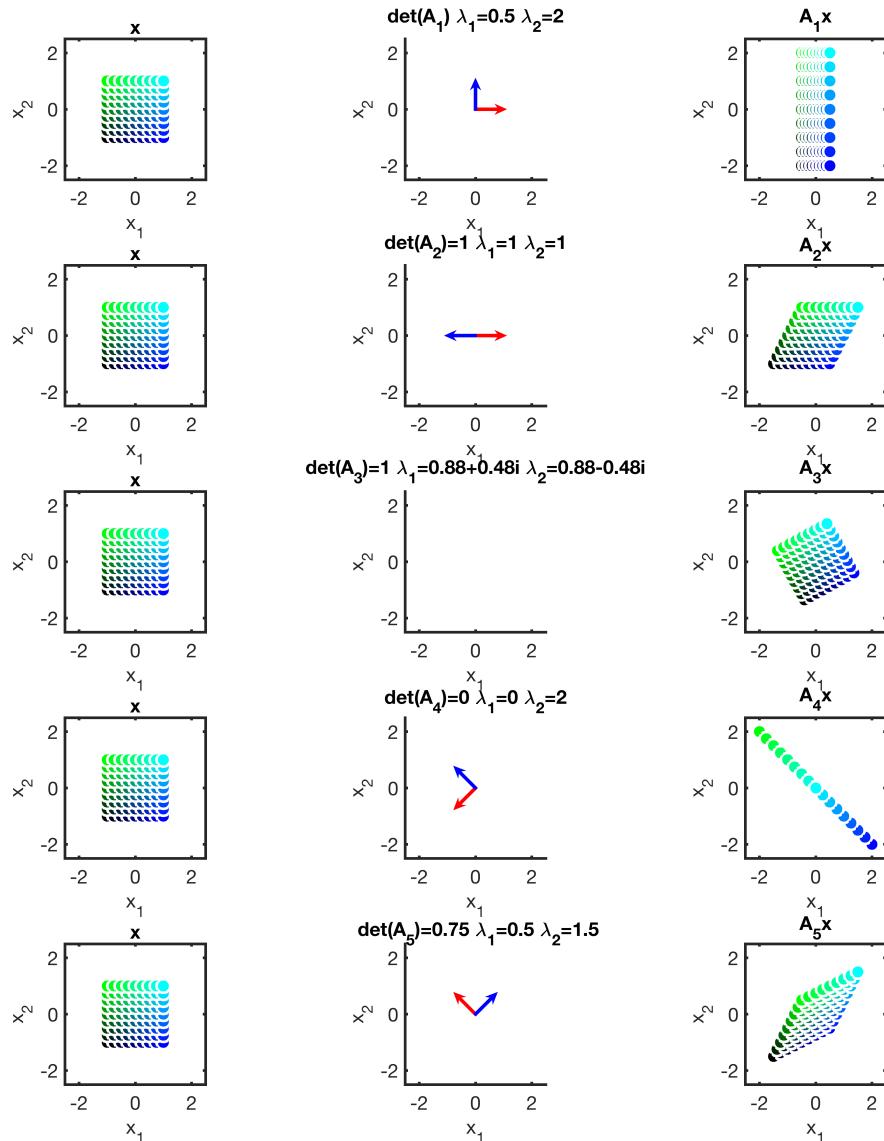
2106 **Theorem 4.12** (Meyer (2000)). *Any symmetric matrix $\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{n \times n}$ has n independent eigenvectors that form an orthogonal basis for \mathbb{R}^n .*

2108 Graphical Intuition in Two Dimensions

2109 Let us gain some intuition for determinants, eigenvectors, eigenvalues and
 2110 how linear maps affect space. Figure 4.4 depicts five transformation matri-
 2111 ces and their impact on a square grid of points. The square grid of points
 2112 are contained within a box of dimensions 2×2 with its centre at the origin.

- 2113 • $\mathbf{A}_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$. The direction of the two eigenvectors correspond to the
 2114 canonical basis vectors in \mathbb{R}^2 , i.e. to two cardinal axes. The horizontal
 2115 axis is compressed by factor $\frac{1}{2}$ (eigenvalue $\lambda_1 = \frac{1}{2}$) and the vertical axis

Figure 4.4
 Determinants and eigenspaces.
 Overview of five linear mappings and their associated transformation matrices
 $A_i \in \mathbb{R}^{2 \times 2}$ project 81 color-coded points $x \in \mathbb{R}^2$ (left column of plots) to target points $A_i x$ (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
 $A_i, i = \{1, \dots, 5\}$.



is extended by a factor of 2 (eigenvalue $\lambda_2 = 2$). The mapping is area preserving ($\det(A_1) = 1 = 2 \times \frac{1}{2}$). Note, that while the area covered by the box of points remained the same, the circumference around the box has increased by 20%.

- $A_2 = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$ 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 ($\det(A_2) = 1$). The eigenvalue $\lambda_1 = 1 = \lambda$)2 is repeated and the hence the eigenvectors are co-linear (drawn here for emphasis

in two opposite directions). This indicating 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 while the mapping is area preserving the circumference around the box of points has increased.

- $A_3 = \begin{bmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$ The rotation matrix 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 3.14 in the corresponding section on rotations.

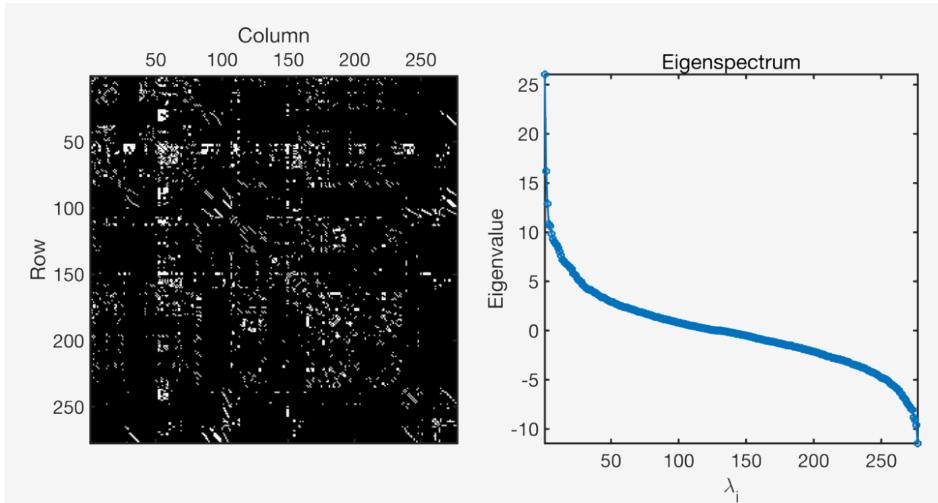
- $A_4 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ reflects a mapping in the standard basis that collapses a two-dimensional domain onto a one-dimensional image space, hence the area is 0. We can see this because one eigenvalue is 0, collapsing the space in direction of the (red) eigenvector corresponding 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%.

- $A_5 = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ is a shear-and-stretch mapping that shrinks space by 75% ($|\det(A_5)| = \frac{3}{4}$), stretching space along the (blue) eigenvector of λ_2 by 50% and compressing it along the orthogonal (red) eigenvector by a factor of 50%.

Example 4.5 (Eigenspectrum of a biological neural network)

Figure 4.5

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



Methods to analyze and learn from network data are an essential component of machine learning methods. 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 \mathbf{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 $\mathbf{A}_{sym} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$. This new matrix \mathbf{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 \mathbf{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.

algebraic
multiplicity

2152 **Definition 4.13.** Let a square matrix \mathbf{A} have an eigenvalue λ_i . The *algebraic multiplicity* of λ_i is the number of times the root appears in the characteristic polynomial.
2153

geometric
multiplicity

2155 **Definition 4.14.** Let a square matrix \mathbf{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 .
2156
2157
2158

2159 *Remark.* A specific eigenvalue's geometric multiplicity must be at least
2160 one, as by definition every eigenvalue has at least one associated eigen-
2161 vector. An eigenvalue's geometric multiplicity cannot exceed its algebraic
2162 multiplicity, but it may be lower. \diamond

Example 4.6

The matrix $A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$ has two repeated eigenvalues $\lambda_1 = \lambda_2 = 2$ and an algebraic multiplicity of 2. The eigenvalue has however only one distinct eigenvector $x_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and thus geometric multiplicity 1.

2163 Before we conclude our considerations of eigenvalues and eigenvectors
2164 it is useful to tie these matrix characteristics together with the previously
2165 covered concept of the determinant and the trace.

Theorem 4.15. *The determinant of a matrix $A \in \mathbb{R}^{n \times n}$ is the product of its eigenvalues, i.e.,*

$$\det(A) = \prod_{i=1}^n \lambda_i, \quad (4.41)$$

2166 where λ_i are (possibly repeated) eigenvalues of A .

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

$$\text{tr}(A) = \sum_{i=1}^n \lambda_i, \quad (4.42)$$

2167 where λ_i are (possibly repeated) eigenvalues of A .

2168 While we leave these two theorems without a proof, we point to the
2169 application of the determinant and trace of the characteristic polynomial
2170 as a way to derive them.

2171 *Remark.* A geometric intuition for these two theorems goes as follows
2172 (see also Figure 4.2 and corresponding text for other examples): Imagine
2173 a unit cube (a box with equal sides of length 1) in \mathbb{R}^3 . We then map the
2174 8 corner points of this box through our matrix A and obtain a new box,
2175 defined by the mapped 8 new corner points. We know that the eigenval-
2176 ues capture the scaling of the basis with respect to the standard basis.
2177 Thus, they capture how the volume of the unit cube (which has volume 1)
2178 was transformed into our box. Thus, the determinant as product of eigen-
2179 values is akin to the volume of the box, a large determinant suggests a
2180 large expansion of volume and vice versa. In contrast the trace is a sum of
2181 eigenvalues, i.e. a sum of length scales. Consider a gift ribbon we would
2182 want to tie around the box. The length of ribbon is proportional to the

2183 length of the sides of the box. The trace of \mathbf{A} captures therefore a notion
2184 of how the matrix acts on the circumference of a volume. \diamond

Example 4.7 (Google's PageRank – Webpages as Eigenvectors)

Google uses the eigenvector corresponding to the maximal eigenvalue of a matrix \mathbf{A} to determine the rank of a page for search. The idea that the PageRank algorithm, developed at Stanford University by Larry Page and Sergey Brin in 1996, came up was that the importance of any web page can be judged by looking at the 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) $x_i \geq 0$ of a website a_i by counting the number of pages pointing to a_i . PageRank also takes the importance of the website into account that links to a website to a_i . Then, the navigation behavior of a user can be described by a transition matrix \mathbf{A} of this graph that tells us with what (click) probability somebody will end up on a different website. The matrix \mathbf{A} has the property that for any initial rank/importance vector \mathbf{x} of a website the sequence $\mathbf{x}, \mathbf{Ax}, \mathbf{A}^2\mathbf{x}, \dots$ converges to a vector \mathbf{x}^* . This vector is called the *PageRank* and satisfies $\mathbf{Ax}^* = \mathbf{x}^*$, i.e., it is an eigenvector (with corresponding eigenvalue 1) of \mathbf{A} . After normalizing by \mathbf{x}^* , such that $\|\mathbf{x}^*\| = 1$ 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

2185

4.3 Cholesky Decomposition

2186 There are many ways to factorize special types of matrices that we en-
2187 counter often in machine learning. In the positive real numbers we have
2188 the square-root operation that yields us a decomposition of the number
2189 into components, for example, $9 = 3 \cdot 3$. For matrices, we need to be
2190 careful that we compute a square-root like operation on positive quanti-
2191 ties. For symmetric, positive definite matrices (see Section 3.2.3) we can
2192 choose from a number of square-root equivalent operations. The *Cholesky*
2193 *decomposition* or *Cholesky factorization* provides a square-root equivalent
2194 operations that is very useful.

Cholesky
decomposition
Cholesky
factorization

Theorem 4.17. *Cholesky Decomposition:* A symmetric positive definite matrix \mathbf{A} can be factorized into a product $\mathbf{A} = \mathbf{LL}^\top$, where \mathbf{L} is a lower triangular matrix with positive diagonal elements:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix}. \quad (4.43)$$

Cholesky factor 2195 \mathbf{L} is called the *Cholesky factor* of \mathbf{A} .

Example 4.8

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.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \equiv \mathbf{L}\mathbf{L}^\top = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} \quad (4.44)$$

Expanding the right hand side yields

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

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

$$l_{11} = \sqrt{a_{11}}, \quad l_{22} = \sqrt{a_{22} - l_{21}^2}, \quad l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)}. \quad (4.45)$$

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

$$l_{21} = \frac{1}{l_{11}}a_{21}, \quad l_{31} = \frac{1}{l_{11}}a_{31}, \quad l_{32} = \frac{1}{l_{22}}(a_{32} - l_{31}l_{21}). \quad (4.46)$$

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 \mathbf{L} should be, given the values a_{ij} for \mathbf{A} and previously computed values of l_{ij} .

2196 The Cholesky decomposition is an important tool for the numerical
2197 computations underlying machine learning. The Cholesky decomposition
2198 is used as a computationally more efficient and numerically more stable
2199 way to solve systems of equations that form symmetric positive definite
2200 matrices, than computing the inverse of such a matrix, and is thus used
2201 under the hood in numerical linear algebra packages.

2202 For matrices that are symmetric positive definite such as the covariance
2203 of a multivariate Gaussian 6.6, one approach is to transform the
2204 matrix into a set of upper or lower triangular matrices. After applying the
2205 Cholesky decomposition we efficiently compute the inverse \mathbf{L}^{-1} of a triangular
2206 matrix by back substitution. Then the original matrix inverse is
2207 computed simply by multiplying the two inverses as $\mathbf{A}^{-1} = (\mathbf{L}\mathbf{L}^\top)^{-1} =$
2208 $(\mathbf{L}^{-1})^\top(\mathbf{L}^{-1})$. As bonus, the determinant is also much easier to compute,
2209 because $\det(\mathbf{A}) = \det(\mathbf{L})^2$, and the determinant of the triangular

2210 Cholesky factor \mathbf{L} is the product of its diagonal elements so that $\det(\mathbf{A}) =$
2211 $\prod_i l_{ii}^2$.

2212 4.4 Eigendecomposition and Diagonalization

Diagonal matrices are of the form

$$\mathbf{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix} \quad (4.47)$$

2213 and possess a very simple structure. Therefore, they allow fast computa-
2214 tion of determinants, powers and inverses. The determinant is the product
2215 of its diagonal entries, a matrix power \mathbf{D}^k is given by each diagonal ele-
2216 ment raised to the power k , and the inverse \mathbf{D}^{-1} is the reciprocal of its
2217 diagonal elements if all of them are non-zero.

2218 In this section, we will look at how to transform matrices into diagonal
2219 form. This is an important application of the basis change we discussed in
2220 Section 2.7.2 and eigenvalues from Section 4.2.

2221 Let us recall that two matrices \mathbf{A}, \mathbf{D} are similar (Definition 2.21) if
2222 there exists an invertible matrix \mathbf{P} , such that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. More specif-
2223 ically, we will look at matrices \mathbf{A} that are similar to a diagonal matrix \mathbf{D}
2224 that contains the eigenvalues of \mathbf{A} on its diagonal.

2225 **diagonalizable** 2226 **Definition 4.18** (Diagonalizable). A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is *diagonalizable*
2227 if it is similar to a diagonal matrix, in other words there exists a matrix
 $\mathbf{P} \in \mathbb{R}^{n \times n}$ so that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$.

2228 In the following, we will see that diagonalizing a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a
2229 way of expressing the same linear mapping but in another basis (see Sec-
2230 tion 2.6.1). Specifically we will try to diagonalize a matrix \mathbf{A} by finding
2231 a new basis that consists of the eigenvectors of \mathbf{A} . We present two theo-
2232 rems, first for square matrices (Theorem 4.19) then for symmetric matri-
2233 ces (Theorem 4.21). The following results parallels the discussion we had
2234 about eigenvalues and eigenvectors (Theorem 4.11 and Theorem 4.12).

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

$$\mathbf{AP} = \mathbf{PD} \quad (4.48)$$

2235 if and only if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of \mathbf{A} and the \mathbf{p}_i are the cor-
2236 responding eigenvectors of \mathbf{A} .

We can see that this statement holds because

$$\mathbf{AP} = \mathbf{A}[\mathbf{p}_1, \dots, \mathbf{p}_n] = [\mathbf{Ap}_1, \dots, \mathbf{Ap}_n] \quad (4.49)$$

$$\mathbf{P}\mathbf{D} = [\mathbf{p}_1, \dots, \mathbf{p}_n] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = [\lambda_1 \mathbf{p}_1, \dots, \lambda_n \mathbf{p}_n] . \quad (4.50)$$

Thus, (4.48) implies that

$$\mathbf{A}\mathbf{p}_1 = \lambda_1 \mathbf{p}_1 \quad (4.51)$$

⋮

$$\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n \quad (4.52)$$

and vice versa.

Thus, the matrix \mathbf{P} must be composed of columns of eigenvectors. But this is not sufficient to know if we can diagonalize \mathbf{A} , as our definition of diagonalization requires that \mathbf{P} is invertible. From Theorem 4.3 we know that our square matrix \mathbf{P} is only invertible (has determinant $\neq 0$) if it has full rank. This implies that the eigenvectors $\mathbf{p}_1, \dots, \mathbf{p}_n$ must be linearly independent. Moreover, consider that Theorem 4.11 tells us when \mathbf{A} is diagonalizable by having n independent eigenvectors, namely in only those cases where \mathbf{A} has n distinct eigenvalues. Taking together these arguments we can now combine them to formulate a key theorem of this chapter.

Theorem 4.19. *Eigendecomposition/Diagonalization theorem.* A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be factored as

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \quad (4.53)$$

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

Diagonalization

Definition 4.20. A defective matrix is a square 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) and is therefore not diagonalizable (see also Theorem 4.11).

defective matrix

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 .

• 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 $S = S^\top \in \mathbb{R}^{n \times n}$ can always be diagonalized into

$$S = PDP^\top \quad (4.54)$$

2265 where P is matrix of n orthogonal eigenvectors and D is a diagonal matrix
2266 of its n eigenvalues.

Proof By Theorem 4.12 we know that $P = [\mathbf{p}_1, \dots, \mathbf{p}_n]$ has n orthogonal eigenvectors of S with eigenvalues $\lambda_1, \dots, \lambda_n$. We can then write

$$(P^\top P)_{ij} = \mathbf{p}_i^\top \mathbf{p}_j \quad (4.55)$$

where

$$\mathbf{p}_i^\top \mathbf{p}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (4.56)$$

2267 and therefore $P^\top P = I$ and $P^{-1} = P^\top$.

We observe the following product

$$\lambda_i P \mathbf{p}_i = \lambda_i [\mathbf{p}_1, \dots, \mathbf{p}_n] \mathbf{p}_i = \lambda_i \mathbf{e}_i, \quad (4.57)$$

which we will use in the following derivation.

$$P^\top S P = P^\top S [\mathbf{p}_1, \dots, \mathbf{p}_n] \quad (4.58)$$

$$= P^\top [S \mathbf{p}_1, \dots, S \mathbf{p}_n] \quad (4.59)$$

$$= P^\top [\lambda_1 \mathbf{p}_1, \dots, \lambda_n \mathbf{p}_n] \quad (4.60)$$

$$= [\mathbf{p}_1, \dots, \mathbf{p}_n]^\top [\lambda_1 \mathbf{p}_1, \dots, \lambda_n \mathbf{p}_n] \quad (4.61)$$

$$= [\lambda_1 \mathbf{e}_1, \dots, \lambda_n \mathbf{e}_n] = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = D \quad (4.62)$$

2268

□

2269 Geometric intuition for the eigendecomposition

2270 We can interpret the eigendecomposition of a matrix as follows (see also
2271 Figure 4.6): Let A be the transformation matrix of a linear mapping with
2272 respect to the standard basis. P^{-1} performs a basis change from the stan-
2273 dard basis into the eigenbasis. This maps the eigenvectors \mathbf{p}_i (red and
2274 green arrows in Figure 4.6) onto the standard axes \mathbf{e}_i . Then, the diagonal
2275 D scales the vectors along these axes by the eigenvalues $\lambda_i \mathbf{e}_i$ and, finally,
2276 P transforms these scaled vectors back into the standard/canonical coor-
2277 dinates (yielding $\lambda_i \mathbf{p}_i$).

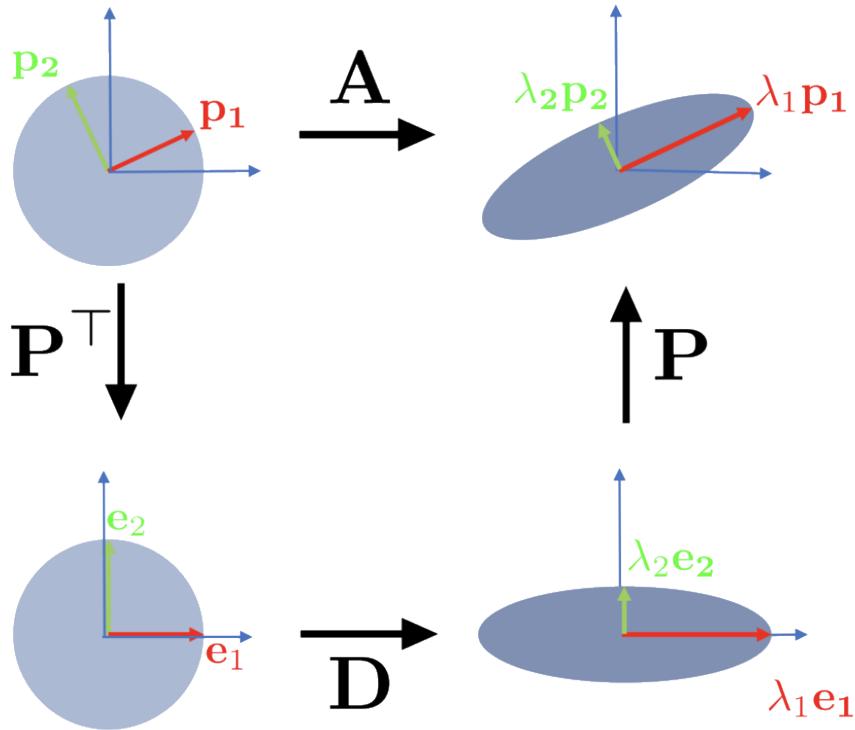


Figure 4.6 Intuition behind the eigendecomposition of a $A \in \mathbb{R}^{2 \times 2}$ in the standard basis as sequential transformations. Top-left to bottom-left: P^\top performs a basis change (here drawn in \mathbb{R}^2 and depicted as a rotation-like operation) mapping the eigenvectors into the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

Example 4.9

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

Step 1: Compute the eigenvalues and eigenvectors

The matrix has eigenvalues

$$\det(A - \lambda I) = \det \left(\begin{bmatrix} 2 - \lambda & 1 \\ 1 & 2 - \lambda \end{bmatrix} \right) \quad (4.63)$$

$$= (2 - \lambda)^2 - 1 = \lambda^2 - 2\lambda + 3$$

$$= (\lambda - 3)(\lambda - 1) = 0. \quad (4.64)$$

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

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} p_1 = 1p_1 \quad (4.65)$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} p_2 = 3p_2. \quad (4.66)$$

This yields

$$\mathbf{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (4.67)$$

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: Compute the diagonalizing matrix \mathbf{P}

To compute the diagonalizing matrix we collect these normalized eigenvectors together

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad (4.68)$$

so that we obtain

$$\begin{aligned} \mathbf{AP} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 3 \\ -1 & 3 \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} = \mathbf{PD}. \end{aligned} \quad (4.69)$$

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

$$\mathbf{A} = \mathbf{PAP}^\top \quad (4.70)$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}. \quad (4.71)$$

²²⁷⁸ The eigenvalue decomposition of a matrix has a number of convenient properties

²²⁷⁹ • Diagonal matrices \mathbf{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 \mathbf{A} via the eigenvalue decomposition

$$\mathbf{A}^k = (\mathbf{PDP}^{-1})^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}. \quad (4.72)$$

²²⁸⁰ Computing \mathbf{D}^k is efficient because we apply this operation individually to any diagonal element.

²²⁸¹ • A different property of diagonal matrices is that they can be used to decouple variables. This will be important in probability theory to in-

2284 terpret random variables, e.g., for the Gaussian distributions we will
2285 encounter in Section 6.6 and in applications such as dimensionality re-
2286 duction Chapter 10.

2287 The eigenvalue decomposition requires square matrices, and for non-
2288 symmetric square matrices it is not guaranteed that we can transform
2289 them into diagonal form. It would be useful to be able to perform a de-
2290 composition on general matrices. In the next section, we introduce a more
2291 general matrix decomposition technique, the Singular Value Decomposi-
2292 tion.

2293 4.5 Singular Value Decomposition

2294 The Singular Value Decomposition (SVD) of a matrix is a central matrix
2295 decomposition method in linear algebra. It has been referred to as the
2296 “fundamental theorem of linear algebra” (Strang, 1993) because it can be
2297 applied to all matrices, not only to square matrices, and it always exists.
2298 Moreover, as we will explore in the following, the SVD of a linear map-
2299 ping $\Phi : V \rightarrow W$ quantifies the resulting change between the underlying
2300 geometry of these two vector spaces. We recommend Kalman (1996); Roy
2301 and Banerjee (2014) for a deeper overview of the mathematics of the SVD.

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

$$\begin{matrix} & n \\ \begin{matrix} A \\ \vdots \end{matrix} & = \end{matrix} \begin{matrix} & m \\ U & \end{matrix} \begin{matrix} & n \\ \Sigma & \end{matrix} \begin{matrix} & n \\ V^\top & \end{matrix} \quad (4.73)$$

Singular Value
Decomposition
SVD

2302 where $U \in \mathbb{R}^{m \times m}$ is an orthogonal matrix composed of column vectors u_i ,
2303 $i = 1, \dots, m$, and $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix of column vectors
2304 v_j , $j = 1, \dots, n$, and Σ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and
2305 $\Sigma_{ij} = 0$, $i \neq j$. The SVD is always possible for any matrix A .

2306 The σ_i are called the singular values, u_i are called the left-singular vec-
2307 tors and v_j are called the right-singular vectors. By convention the singular
2308 vectors are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \sigma_r \geq 0$.

singular values
left-singular vectors
right-singular
vectors

2309 We will see a proof of this theorem later in this section. The SVD al-
2310 lows us to decompose general matrices, and the existence of the unique
2311 singular value matrix Σ requires attention. Observe that the $\Sigma \in \mathbb{R}^{m \times n}$ is
2312 rectangular, that is it is non-square. In particular note that Σ is the same
2313 size as A . This means that Σ has a diagonal submatrix that contains the
2314 singular values and needs additional zero vectors that increase the dimen-
2315 sion.

singular value
matrix

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

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}. \quad (4.74)$$

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

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & & 0 \\ 0 & 0 & \sigma_n & 0 & \dots & 0 \end{bmatrix} \quad (4.75)$$

2316 4.5.1 Geometric Intuitions for the SVD

2317 The SVD has a number of interesting geometric intuitions to offer to de-
 2318 scribe a transformation matrix. Broadly there are two intuitive views we
 2319 can have. First we consider the SVD as sequential operations performed
 2320 on the bases (discussed in the following), and second we consider the
 2321 SVD as operations performed on sets of (data) points as described in Ex-
 2322 ample 4.10.

2323 The SVD can be interpreted as a decomposition of a linear mapping
 2324 (recall Section 2.7.1) $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ into three operations (see Figure 4.7
 2325 for the following). The SVD intuition follows superficially a similar struc-
 2326 ture to our eigendecomposition intuition (confront Figure 4.7 for the SVD
 2327 with Figure 4.6 for the eigendecomposition: Broadly speaking the SVD
 2328 performs a basis change (\mathbf{V}^\top) followed by a scaling and augmentation
 2329 (or reduction) in dimensionality (Σ) and then performs a second basis
 2330 change (\mathbf{U}). The SVD entails a number of important details and caveats
 2331 which is why we will review our intuition in more detail and precision,
 2332 than we have had for the eigendecomposition.

2333 Assume we are given a transformation matrix of Φ with respect to the
 2334 standard bases B and C of \mathbb{R}^n and \mathbb{R}^m , respectively. Moreover, assume a
 2335 second basis \tilde{B} of \mathbb{R}^n and \tilde{C} of \mathbb{R}^m . Then

- 2336 1. \mathbf{V} performs a basis change in the domain \mathbb{R}^n from \tilde{B} (represented
 2337 by the red and green vectors \mathbf{v}_1 and \mathbf{v}_2 in Figure 4.7 top left) to the
 2338 canonical basis B . It is useful here to recall our discussion of basis
 2339 changes Section 2.7.2 and orthogonal matrices and orthonormal bases
 2340 in Section ??), as $\mathbf{V}^\top = \mathbf{V}^{-1}$ performs a basis change from B to \tilde{B}
 2341 (the red and green vectors are now aligned with the canonical basis in
 2342 Figure 4.7 bottom left).

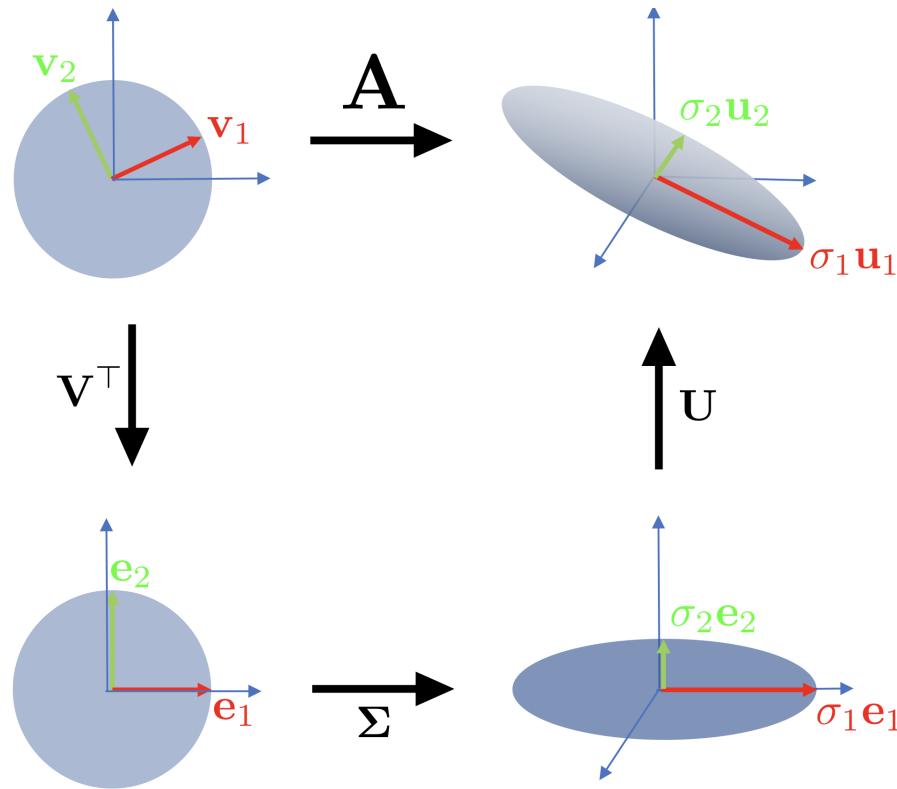
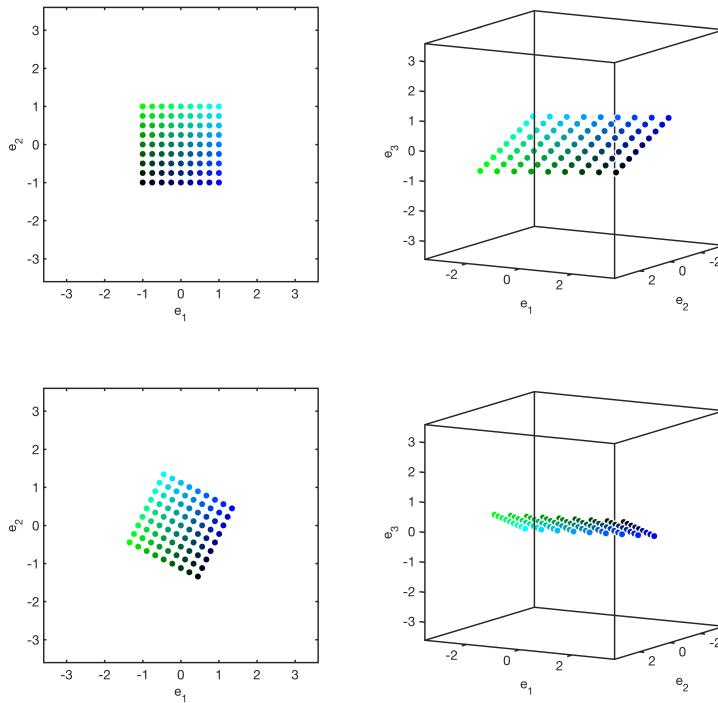


Figure 4.7 Intuition behind SVD of a $A \in \mathbb{R}^{3 \times 2}$ in the standard basis as sequential transformations. Top-left to bottom-left: V^\top performs a basis change in \mathbb{R}^2 . Bottom-left-to-bottom right Σ performs a scaling and increases the dimensionality from \mathbb{R}^2 to \mathbb{R}^3 . 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-left to top-left: U performs a second basis change within \mathbb{R}^3 .

- 2343 2. Having changed the coordinate system to \tilde{B} , Σ scales the new coordinates by the singular values σ_i (and adding or deleting dimensions),
 2344 i.e., Σ 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 e_1 - e_2 plane which is now embedded in a third dimension in Figure 4.7 bottom right).
 2345
 2346
 2347
 2348
 2349 3. 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 e_1 - e_2 plane in Figure 4.7 bottom right).

2352 The SVD expresses a change of basis in both the domain and codomain:
 2353 The columns of U and V are the bases \tilde{B} of \mathbb{R}^n and \tilde{C} of \mathbb{R}^m , respectively.
 2354 Note, how this is in contrast with the eigendecomposition that operates
 2355 within the same vector space (where the same basis change is applied and
 2356 then undone). What makes the SVD special is that these two (different)
 2357 bases are simultaneously linked by the singular values matrix Σ . We refer
 2358 to Section 2.7.2 and Figure 2.9 for a more detailed discussion on basis
 2359 change.

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.



Example 4.10

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 standard basis we map these points using

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (4.76)$$

$$= \mathbf{U} \Sigma \mathbf{V}^\top \quad (4.77)$$

$$= \begin{bmatrix} 0.913 & 0 & -0.408 \\ -0.365 & 0.4472 & -0.816 \\ 0.182 & 0.894 & 0.4082 \end{bmatrix} \begin{bmatrix} 2.449 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4472 & -0.894 \\ 0.8941 & 0.4472 \end{bmatrix} \quad (4.78)$$

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 $\mathbf{V}^\top \in \mathbb{R}^{2 \times 2}$ are shown in the bottom-left panel of Figure 4.8. After a mapping Σ to the codomain \mathbb{R}^3

(see bottom right panel in Figure 4.8) we can see how all the points lie on the e_1 - 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 \mathbf{A} to the codomain \mathbb{R}^3 equals the transformation of \mathcal{X} by $\mathbf{U}\Sigma\mathbf{V}^\top$, where \mathbf{U} performs a rotation within the codomain \mathbb{R}^3 so that the mapped points are no longer restricted to the e_1 - 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 matrix

$$\mathbf{S} = \mathbf{S}^\top = \mathbf{P}\mathbf{D}\mathbf{P}^\top \quad (4.79)$$

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

$$\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^\top. \quad (4.80)$$

We identify

$$\mathbf{U} = \mathbf{P} = \mathbf{V}, \quad (4.81)$$

$$\mathbf{D} = \Sigma, \quad (4.82)$$

so that the SVD of symmetric matrices is their eigenvalue decomposition. \diamond

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

Our plan is to start with constructing the orthonormal set of right-singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$. We then construct the orthonormal set of left-singular vectors $\mathbf{u}_1, \dots, \mathbf{u}_m \in \mathbb{R}^m$. Thereafter, we will link the two and require that the orthogonality of the \mathbf{v}_i is preserved under the transformation of \mathbf{A} . This is important because we know the images \mathbf{Av}_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 4.10 we can always construct a symmetric matrix $\mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n \times n}$ from any rectangular matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. Thus, we can always diagonalize $\mathbf{A}^\top \mathbf{A}$ and obtain

$$\mathbf{A}^\top \mathbf{A} = \mathbf{P} \mathbf{D} \mathbf{P}^\top = \mathbf{P} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \mathbf{P}^\top. \quad (4.83)$$

Take note that the $\lambda_i \geq 0$ are the eigenvalues of $\mathbf{A}^\top \mathbf{A}$. Let us assume the SVD of \mathbf{A} exists and inject (4.73) into (4.83).

$$\mathbf{A}^\top \mathbf{A} = (\mathbf{U} \Sigma \mathbf{V}^\top)^\top (\mathbf{U} \Sigma \mathbf{V}^\top) = \mathbf{V} \Sigma^\top \mathbf{U}^\top \mathbf{U} \Sigma \mathbf{V}^\top. \quad (4.84)$$

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

$$\mathbf{A}^\top \mathbf{A} = \mathbf{V} \Sigma^\top \Sigma \mathbf{V}^\top = \mathbf{V} \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix} \mathbf{V}^\top. \quad (4.85)$$

Comparing now (4.83) and (4.85) we identify

$$\mathbf{V} = \mathbf{P}, \quad (4.86)$$

$$\sigma_i^2 = \lambda_i. \quad (4.87)$$

2380 Therefore, the eigenvectors \mathbf{P} of $\mathbf{A}^\top \mathbf{A}$ are the right-singular vectors \mathbf{V} of 2381 \mathbf{A} (see (4.86)). They form an orthonormal basis because of Theorem 4.21, 2382 for the domain of the SVD. Moreover, the eigenvalues of $\mathbf{A}^\top \mathbf{A}$ are the 2383 squared singular values of Σ (see (4.87)).

Let us now repeat this derivation but this time we will focus on obtaining the left singular vectors \mathbf{U} instead of \mathbf{V} . Therefore we start again by computing the SVD of a symmetric matrix, this time $\mathbf{A} \mathbf{A}^\top \in \mathbb{R}^{m \times m}$ (instead of the above $\mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n \times n}$). We inject again (4.73) and obtain:

$$\mathbf{A} \mathbf{A}^\top = (\mathbf{U} \Sigma \mathbf{V}^\top) (\mathbf{U} \Sigma \mathbf{V}^\top)^\top = \mathbf{U} \Sigma^\top \mathbf{V}^\top \mathbf{V} \Sigma \mathbf{U}^\top \quad (4.88)$$

$$= \mathbf{U} \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m^2 \end{bmatrix} \mathbf{U}^\top. \quad (4.89)$$

2384 We can now obtain from the same arguments about symmetric matrices 2385 and their diagonalization, now applied to $\mathbf{A} \mathbf{A}^\top$, the orthonormal eigen- 2386 vectors of $\mathbf{A}^\top \mathbf{A}$. These are the left-singular vectors \mathbf{U} and form an or- 2387 thonormal basis set in the codomain of the SVD.

This leaves the question of the structure of the matrix Σ . We need to show that regardless of $n > m$ or $n < m$, that $\mathbf{A} \mathbf{A}^\top$ and $\mathbf{A}^\top \mathbf{A}$ have the

same non-zero eigenvalues: Let us assume that λ is a non-zero eigenvalue of $\mathbf{A}\mathbf{A}^\top$ and \mathbf{x} is an eigenvector belonging to λ_i . Then

$$(\mathbf{A}\mathbf{A}^\top)\mathbf{x} = \lambda\mathbf{x} \quad (4.90)$$

left multiplying by \mathbf{A} yields and pulling on the right-hand side the scalar factor λ forward

$$\mathbf{A}(\mathbf{A}\mathbf{A}^\top)\mathbf{x} = \mathbf{A}(\lambda\mathbf{x}) = \lambda(\mathbf{A}\mathbf{x}) \quad (4.91)$$

and we can use (2.30) to reorder the left-hand side factors

$$(\mathbf{A}^\top\mathbf{A})(\mathbf{A}^\top\mathbf{x}) = \lambda(\mathbf{A}\mathbf{x}). \quad (4.92)$$

2388 This is the eigenvalue equation for $\mathbf{A}\mathbf{A}^\top$. Therefore, λ is the same eigenvalue 2389 for $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$, and $\mathbf{A}\mathbf{x}$ is its eigenvector. Thus, both matrices 2390 have the same non-zero eigenvalues. Thus, the Σ matrices in the SVD for 2391 both cases have to be the same.

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

$$(\mathbf{Av}_i)^\top(\mathbf{Av}_j) = \mathbf{v}_i^\top(\mathbf{A}^\top\mathbf{A})\mathbf{v}_j = \mathbf{v}_i^\top(\lambda_j\mathbf{v}_j) = \lambda_j\mathbf{v}_i^\top\mathbf{v}_j = 0. \quad (4.93)$$

2392 For the case $m > r$ this holds for all pairs $\mathbf{Av}_1, \dots, \mathbf{Av}_r$ the images are 2393 a basis of \mathbb{R}^m , while if any further vectors $\mathbf{Av}_i, i > r$ exist, they must be 2394 in the nullspace of \mathbf{A} (see remark after proof for the converse case).

To complete the SVD construction we need left-singular vectors that are orthonormal: we normalize the images of the right-singular vectors \mathbf{Av}_i and call them \mathbf{u}_i ,

$$\mathbf{u}_i = \frac{\mathbf{Av}_i}{\|\mathbf{Av}_i\|} = \frac{1}{\sqrt{\lambda_i}}\mathbf{Av}_i = \frac{1}{\sigma_i}\mathbf{Av}_i \quad (4.94)$$

2395 where the last equality was obtained from (4.87) and from equation (4.89) 2396 showing us that the eigenvalues of $\mathbf{A}\mathbf{A}^\top$ are such that $\sigma_i^2 = \lambda_i$.

2397 Therefore, the eigenvectors of $\mathbf{A}^\top\mathbf{A}$, which we know are the right- 2398 singular vectors \mathbf{v}_i and their normalized images under \mathbf{A} , the left singular 2399 vectors \mathbf{u}_i , form two self-consistent sets of orthonormal bases that are coupled by the singular value matrix Σ .

Remark. Let us rearrange (4.94) to obtain the *singular value equation*

$$\mathbf{Av}_i = \sigma_i\mathbf{u}_i, \quad i = 1, \dots, r. \quad (4.95)$$

singular value
equation

2401 This equation closely resembles the eigenvalue equation (4.27), but the 2402 vectors on the left and the right-hand sides are not the same.

For $n > m$ (4.95) holds only for $i \leq m$ and (4.95) say nothing about the \mathbf{u}_i for $i > m$, but we know by construction that they are orthonormal. Conversely for $m > n$, then (4.95) holds only for $i \leq n$. For $i > n$ we have $\mathbf{A}\mathbf{v}_i = 0$ and we still know that the \mathbf{v}_i form an orthonormal set. This means that the SVD also supplies an orthonormal basis of the kernel (or null space) or \mathbf{A} , the set of vectors \mathbf{x} with $\mathbf{Ax} = 0$ (see Section 2.7.3).

Moreover, collecting the \mathbf{v}_i as the columns of \mathbf{V} and \mathbf{u}_i as the columns of \mathbf{U} yields

$$\mathbf{AV} = \mathbf{U}\Sigma. \quad (4.96)$$

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

Example 4.11

Let us find the singular value decomposition of

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}. \quad (4.97)$$

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

$$\mathbf{A}^\top \mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 2 \\ 1 & 0 & 1 \end{bmatrix}. \quad (4.98)$$

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

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

$$\mathbf{A}^\top \mathbf{A} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad (4.99)$$

$$= \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{2}} \\ \frac{1}{\sqrt{30}} & \frac{\sqrt{5}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{2}} & \frac{-2}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \mathbf{P}\mathbf{D}\mathbf{P}^\top. \quad (4.100)$$

Note, that due to our orthonormality requirement implies that we chose the 3rd column of \mathbf{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 $\mathbf{A}^\top \mathbf{A}$ we obtain them straight from \mathbf{D} . Note that because $\text{rk}(\mathbf{A}) = 2$ 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 \mathbf{A} , hence,

$$\Sigma = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \quad (4.101)$$

We also have obtained already the right-singular vectors because

$$\mathbf{V} = \mathbf{P} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{\sqrt{2}}{2} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \quad (4.102)$$

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 \mathbf{A} and normalizing them by dividing them by their corresponding singular value.

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}, \quad (4.103)$$

$$\mathbf{u}_2 = \frac{1}{\sigma_2} \mathbf{A} \mathbf{v}_2 = \frac{1}{1} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} \quad (4.104)$$

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2] = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}. \quad (4.105)$$

Note that in practice the approach illustrated here has poor numerical behaviour, and the SVD of \mathbf{A} is computed without resorting to the eigenvalue decomposition of $\mathbf{A}^\top \mathbf{A}$.

4.5.3 Eigenvalue Decomposition vs Singular Value Decomposition

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

The SVD always exists for any matrix $\mathbb{R}^{n \times m}$. 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 \mathbf{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 \mathbf{U} and \mathbf{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.

$$\begin{array}{c}
 \begin{array}{ccccc}
 & \text{Ali} & \text{Beatrix} & \text{Chandra} & \\
 \begin{matrix}
 \text{Star Wars} \\ \text{Blade Runner} \\ \text{Amelie} \\ \text{Delicatessen}
 \end{matrix} &
 \left[\begin{array}{cccc}
 5 & 4 & 1 & \\
 5 & 5 & 0 & \\
 0 & 0 & 5 & \\
 1 & 0 & 4 &
 \end{array} \right] = &
 \left[\begin{array}{cccc}
 -0.6710 & 0.0236 & 0.4647 & -0.5774 \\
 -0.7197 & 0.2054 & -0.4759 & 0.4619 \\
 -0.0939 & -0.7705 & -0.5268 & -0.3464 \\
 -0.1515 & -0.6030 & 0.5293 & 0.5774
 \end{array} \right] & \\
 & & \left[\begin{array}{ccc}
 9.6438 & 0 & 0 \\
 0 & 6.3639 & 0 \\
 0 & 0 & 0.7056 \\
 0 & 0 & 0
 \end{array} \right] & \\
 & & & \left[\begin{array}{ccc}
 -0.7367 & -0.6515 & -0.1811 \\
 0.0852 & 0.1762 & -0.9807 \\
 0.6708 & -0.7379 & -0.0743
 \end{array} \right]
 \end{array}
 \end{array}$$

2428 A key difference between the eigendecomposition and the SVD is that
 2429 in the SVD, domain and co-domain can be vector spaces of different
 2430 dimensions.

- 2431 • In the SVD, the left and right singular vector matrices \mathbf{U} and \mathbf{V} are generally not inverse of each other. In the eigendecomposition the eigen-
 2432 vector matrices \mathbf{P} and \mathbf{P}^{-1} are inverses of each other.
- 2433 • In the SVD, the entries in the diagonal matrix Σ are all real and nonneg-
 2434 ative, which is not generally true for the diagonal matrix in the eigen-
 2435 decomposition.
- 2436 • The SVD and the eigendecomposition are closely related through their
 2437 projections
 - 2438 – The left-singular vectors of \mathbf{A} are eigenvectors of $\mathbf{A}\mathbf{A}^\top$
 - 2439 – The right-singular vectors of \mathbf{A} are eigenvectors of $\mathbf{A}^\top\mathbf{A}$.
 - 2440 – The non-zero singular values of \mathbf{A} are the square roots of the non-
 2441 zero eigenvalues of $\mathbf{A}\mathbf{A}^\top$, and equal the non-zero eigenvalues of
 2442 $\mathbf{A}^\top\mathbf{A}$.
- 2443 • For symmetric matrices the eigenvalue decomposition and the SVD are
 2444 one and the same.

Example 4.12 (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 $\mathbf{A} \in \mathbb{R}^{4 \times 3}$ (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 \mathbf{x}_{Ali} , $\mathbf{x}_{\text{Beatrix}}$, $\mathbf{x}_{\text{Chandra}}$.

Factoring \mathbf{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 \mathbf{u}_i as stereotypical movies and the right-singular vectors \mathbf{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 \mathbf{v}_j . Similarly, any movie's like-ability can be expressed as a linear combination of the \mathbf{u}_i .

Let us look at the specific outcome of performing SVD: The first left-singular vector \mathbf{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 \mathbf{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 \mathbf{v}_1 may reflect an idealized notion of a science fiction lover.

Similarly, \mathbf{u}_2 , seems to capture a French art house film theme, and \mathbf{v}_2 may be reflecting that Chandra is to 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 \mathbf{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.

2446 2447 2448 2449 *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:

- 2450 2451 2452 2453
- 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 (4.73) for the SVD is sometimes called the *full SVD*.
 - Some authors define the SVD a bit differently, for $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $m \geq n$

full SVD

$$\mathbf{A}_{m \times n} = \mathbf{U}_{m \times n, n \times n, n \times n} \Sigma \mathbf{V}^T \quad (4.106)$$

2454 2455 2456 2457 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

reduced SVD

2458 alternative notation is that Σ is diagonal, as in the eigenvalue decom-
 2459 position. However, it loses the interpretation of Σ as a transformation
 2460 matrix.

- 2461 • In Section 4.6, we will learn about matrix approximation techniques
 2462 using the SVD, which is also called the *truncated SVD*.
- 2463 • One can also define the SVD of a rank- r matrix A so that U is an
 2464 $m \times r$ matrix, Σ as a diagonal matrix $r \times r$, and V as $r \times n$ matrix.
 2465 This construction is very similar to our definition, and ensures that the
 2466 diagonal matrix Σ has only non-zero entries along the diagonal. The
 2467 main convenience of this alternative notation is that Σ is diagonal, as
 2468 in the eigenvalue decomposition.
- 2469 • One could also introduce the restriction that the SVD for A only applies
 2470 to $m \times n$ matrices with $m > n$. However, this restriction is practically
 2471 unnecessary. When $m < n$ the SVD decomposition will yield Σ with
 2472 more zero columns than rows and, consequently, the singular values
 2473 $\sigma_{m+1}, \dots, \sigma_n$ are implicitly 0.

◊

2474 The SVD is used in a variety of applications in machine learning from
 2475 least squares problems in curve fitting to solving systems of linear equa-
 2476 tions. These applications harness various important properties of the SVD,
 2477 its relation to the rank of a matrix and its ability to approximate matrices
 2478 of a given rank with lower rank matrices. Substituting the SVD form of a
 2479 matrix in computations rather use the original matrix has often the advan-
 2480 tage of making the calculation more robust to numerical rounding errors.
 2481 As we will explore in the next section the SVD's ability to approximate
 2482 matrices with "simpler" matrices in a principled manner opens up ma-
 2483 chine learning applications ranging from dimensionality reduction, topic
 2484 modeling to data compression and clustering.

2486 4.6 Matrix Approximation

2487 We will now investigate how the SVD allows us to represent a matrix A
 2488 as a sum of simpler matrices A_i .

Let us construct a rank-1 $m \times n$ matrix A_i as

$$2489 A_i = u_i v_i^\top \quad (4.107)$$

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

$$2490 A = \sum_{i=1}^r \sigma_i u_i v_i^\top = \sum_{i=1}^r \sigma_i A_i \quad (4.108)$$

2491 where the outer product matrices A_i are weighed by the size of the i th

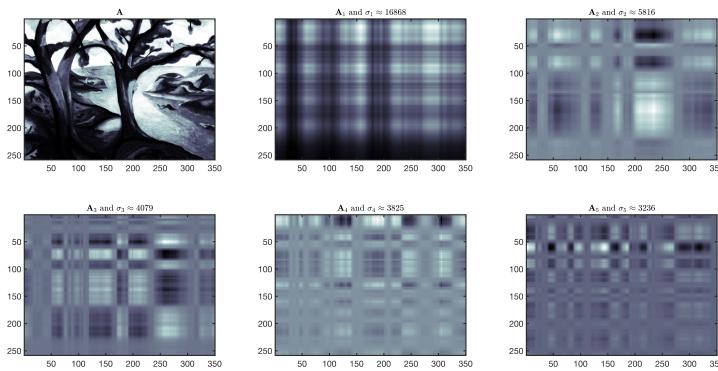


Figure 4.10 (Top left) A grayscale image is a 280×350 matrix of values between 0 (black) and 1 (white). (Middle left to Bottom right) rank-1 matrices $\mathbf{A}_1 \dots \mathbf{A}_5$ and their corresponding singular values $\sigma_1, \dots, \sigma_5$. Note, that the grid like structure of each rank-1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

2490 singular value σ_i . Thus, the sum of the outer products of matching left
 2491 and right singular vectors (weighted by their singular value) is equal to
 2492 \mathbf{A} . Note, that any terms $i > r$ are zero, as the singular values will be
 2493 0. We can see why (4.107) holds: the diagonal structure of the singular
 2494 value matrix Σ multiplies only matching left- and right-singular vectors
 2495 $(\mathbf{u}_i, \mathbf{v}_i^\top)$ and adds them up, while setting non-matching left- and right-
 2496 singular vectors $(\mathbf{u}_i, \mathbf{v}_j^\top, i \neq j)$ to zero.

In the previous paragraph we introduced a low-rank matrix \mathbf{A}_i (of rank 1). We summed up the r individual rank-1 matrices to obtain a rank r matrix \mathbf{A} . What happens if the sum does not over all matrices \mathbf{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 \mathbf{A} that we call the *rank-k approximation* $\widehat{\mathbf{A}}(k)$

$$\widehat{\mathbf{A}}(k) = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad (4.109)$$

2497 of \mathbf{A} with $\text{rk}(\widehat{\mathbf{A}}) = k$.

2498 It would be useful if we could measure how large the difference be-
 2499 tween \mathbf{A} and its approximation $\widehat{\mathbf{A}}(k)$ is in terms of a single number – we
 2500 thus need the notion of a norm. We have already used norms on vectors
 2501 that measure the length of a vector. By analogy we can also define a norm
 2502 on matrices (one of the many ways to define matrix norms).

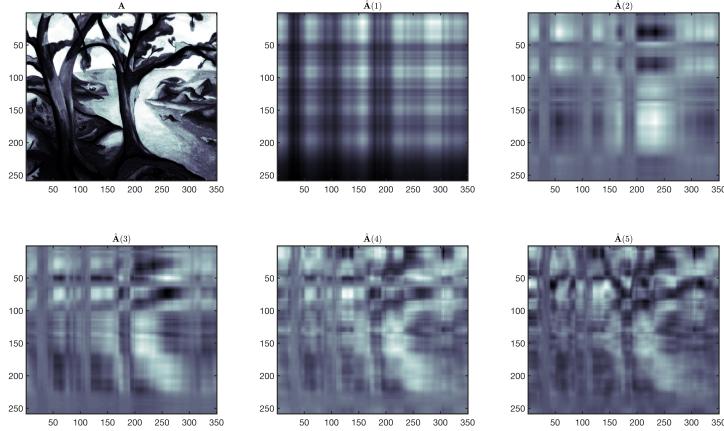
Definition 4.23 (Spectral norm of a matrix). The spectral norm of a ma-
 trix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as the following for $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{A}\|_2 := \max_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \quad \mathbf{x} \neq \mathbf{0}. \quad (4.110)$$

2503 The operator norm implies how long any vector \mathbf{x} can at most become
 2504 once it is multiplied by \mathbf{A} . This maximum lengthening is given by the SVD
 2505 of \mathbf{A} .

rank-k
approximation

Figure 4.11 (Top left) The same grayscale image as in Figure 4.10. (Middle left to Bottom right) Image reconstruction using the low-rank approximation of the SVD: (Top middle) is $\widehat{\mathbf{A}}(1) = \sigma_1 \mathbf{A}_1$. (Top right) is the rank-2 approximation $\widehat{\mathbf{A}}(2) = \sigma_1 \mathbf{A}_1 + \sigma_2 \mathbf{A}_2$. (Bottom left to Bottom right) are $\widehat{\mathbf{A}}(3)$ to $\widehat{\mathbf{A}}(5)$. Note how the shape of the trees becomes²⁵⁰⁶ increasingly visible and clearly recognizable in the a rank-6 approximation. While the original image requires $280 \times 350 = 98000$ numbers, the rank-6 approximation requires us only to store only the 6 singular values and the 6 left and right singular vectors (255 and 380 dimensional each) for a total of $6 \times (250+380+1) = 3786$ numbers – just about 4% of the original.



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

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

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} = \max_{\mathbf{x}} \sqrt{\frac{\|\mathbf{Ax}\|_2^2}{\|\mathbf{x}\|_2^2}} \quad (4.111)$$

$$= \max_{\mathbf{x}} \sqrt{\frac{(\mathbf{x}\mathbf{A})^\top(\mathbf{Ax})}{\mathbf{x}^\top\mathbf{x}}} = \max_{\mathbf{x}} \sqrt{\frac{\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}} \quad (4.112)$$

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

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x}} \sqrt{\frac{\mathbf{x}^\top(\mathbf{P}\mathbf{D}\mathbf{P}^\top)\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}}, \quad (4.113)$$

$$(4.114)$$

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

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_1} \quad (4.115)$$

the largest eigenvalue of $\mathbf{A}^\top\mathbf{A}$ is by (4.87) the largest singular value of \mathbf{A}

$$\|\mathbf{A}\|_2 = \sigma_1 \quad (4.116)$$

Theorem 4.25 (Eckart-Young (or Eckart-Young-Minsky) theorem)). Let

$\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix of rank r and $\mathbf{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k . For any $k \leq r$ such that $\widehat{\mathbf{A}}(k) = \sum_i^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, it holds that

$$\|\mathbf{A} - \widehat{\mathbf{A}}(k)\|_2 = \sigma_{k+1} \quad (4.117)$$

$$= \min_{\text{rk}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2. \quad (4.118)$$

2507 Remark. We can interpret the rank- k approximation obtained with the
 2508 SVD as a projection of the full rank matrix \mathbf{A} onto the lower-dimensional
 2509 space of rank at-most- k matrices. Of all possible projections the SVD rank-
 2510 k approximation minimizes the difference with respect to the spectral
 2511 norm between \mathbf{A} and any rank- k matrix. \diamond

We can retrace some of the steps to understand why (4.117) should hold. We observe that the difference between $\mathbf{A} - \widehat{\mathbf{A}}(k)$ is a matrix containing the sum of the remaining rank-1 matrices

$$\mathbf{A} - \widehat{\mathbf{A}}(k) = \sum_{i=k+1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad (4.119)$$

2512 Thus, by applying the definition of the spectral norm, (4.110), the most
 2513 a vector can be lengthened by the difference matrix is given its largest
 2514 singular value i.e. σ_{k+1} , which is the difference matrix's spectral norm.

Let us proceed to better understand (4.118) validity. We assume that there is another matrix \mathbf{B} with $\text{rk}(\mathbf{B}) \leq k$ such that

$$\|\mathbf{A} - \mathbf{B}\|_2 < \|\mathbf{A} - \widehat{\mathbf{A}}(k)\|_2 \quad (4.120)$$

Then there exists an $(n - k)$ -dimensional nullspace $Z \subseteq \mathbb{R}^n$ such that $\mathbf{x} \in Z \implies \mathbf{Bx} = \mathbf{0}$. In other words, have an n -dimensional space \mathbb{R}^n in which lies a lower dimensional nullspace of \mathbf{B} . Then it follows that

$$\|\mathbf{Ax}\|_2 = \|(\mathbf{A} - \mathbf{B})\mathbf{x}\|_2, \quad (4.121)$$

and by using a version of the Cauchy-Schwartz inequality (3.5) that encompasses norms of matrices we obtain

$$\|\mathbf{Ax}\|_2 \leq \|\mathbf{A} - \mathbf{B}\|_2 \|\mathbf{x}\|_2 < \sigma_{k+1} \|\mathbf{x}\|_2 \quad (4.122)$$

2515 Therefore, V is a $(n - k)$ dimensional subspace where $\|\mathbf{Ax}\|_2 < \sigma_{k+1} \|\mathbf{x}\|_2$.

2516 On the other hand there is a $(n + 1)$ -dimensional subspace where $\|\mathbf{Ax}\|_2 \geq$
 2517 $\sigma_{k+1} \|\mathbf{x}\|_2$ which is spanned by the right singular vector \mathbf{v}_{k+1} of \mathbf{A} . Adding
 2518 up dimensions of these two spaces yields a number greater n , as there
 2519 must be a non-zero vector in both spaces. This is a contradiction because
 2520 of the Rank-Nullity Theorem (recall Theorem 2.23 in Section 2.7.3).

2521 The Eckart-Young theorem implies that we can use SVD to reduce a
 2522 rank- r matrix \mathbf{A} to a rank- k matrix $\widehat{\mathbf{A}}$ in a principled, optimal (in the
 2523 spectral norm sense) manner. The effect of the low-rank approximation
 2524 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 4.13 (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

$$\mathbf{M}_1 = \sigma_1(\mathbf{u}_1 \mathbf{v}_1^\top) \quad (4.123)$$

$$= 9.6438 \begin{bmatrix} -0.6710 \\ -0.7197 \\ -0.0939 \\ -0.1515 \end{bmatrix} \begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \end{bmatrix} \quad (4.124)$$

$$= \begin{bmatrix} 4.7673 & 4.2154 & 1.1718 \\ 5.1138 & 4.5218 & 1.2570 \\ 0.6671 & 0.5899 & 0.1640 \\ 1.0765 & 0.9519 & 0.2646 \end{bmatrix} \quad (4.125)$$

This first rank-1 approximation \mathbf{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.

$$\mathbf{M}_2 = \sigma_2(\mathbf{u}_2 \mathbf{v}_2^\top) \quad (4.126)$$

$$= 6.3639 \begin{bmatrix} 0.0236 \\ 0.2054 \\ -0.7705 \\ -0.6030 \end{bmatrix} \begin{bmatrix} 0.0852 & 0.1762 & -0.9807 \end{bmatrix} \quad (4.127)$$

$$= \begin{bmatrix} 0.0128 & 0.0265 & -0.1475 \\ 0.1114 & 0.2304 & -1.2820 \\ -0.4178 & -0.8642 & 4.8084 \\ -0.3270 & -0.6763 & 3.7631 \end{bmatrix} \quad (4.128)$$

In this second rank-1 approximation \mathbf{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 $\hat{\mathbf{A}}(2)$ where we combine the first two rank-1 approximations

$$\hat{\mathbf{A}}(2) = \mathbf{M}_1 + \mathbf{M}_2 \quad (4.129)$$

$$= \begin{bmatrix} 4.7801 & 4.2419 & 1.0244 \\ 5.2252 & 4.7522 & -0.0250 \\ 0.2493 & -0.2743 & 4.9724 \\ 0.7495 & 0.2756 & 4.0278 \end{bmatrix} \quad (4.130)$$

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

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} \quad (4.131)$$

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 spanned in our example by a two-dimensional space spanned by science fiction and French art house movies and lovers.

2531

4.7 Matrix Phylogeny

2532 In Chapter 2 and 3 we covered the basics of linear algebra and analytic
 2533 geometry, in this chapter we now looked at fundamental characteristics
 2534 and methods on matrices and linear mappings. We are depicting in Fig-
 2535 ure 4.12 the phylogenetic tree of relationships between different types of
 2536 matrices (black arrows indicating “is a subset of”) and the covered opera-
 2537 tions we can perform on them (in red). For example, we already learned
 2538 in Chapter 2 about **square** matrices, which are a subset of **all (complex)**
 2539 **matrices** (top level node in the tree). We will then learn here that we can
 2540 compute a specific characteristic (**determinant**) in Section 4.1 that will
 2541 inform us whether a square matrix has an associate **inverse matrix**, thus
 2542 if it belongs to the class of non-singular, invertible matrices.

2543 Going backward through the chapter, we start with the most general
 2544 case of real matrices $\mathbb{R}^{n \times m}$ for which we can define a pseudo-inverse to
 2545 “invert” them, as well as perform **singular value decomposition (SVD)**
 2546 (Theorem 4.22). This superset of matrices is divided into the square $\mathbb{R}^{n \times n}$
 2547 matrices for which we can define the characteristic feature of the **deter-
 2548 minant** and the **trace** (Section 4.1).

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

2549 Here the set of matrices splits in two: If the square $\mathbb{R}^{n \times n}$ matrix has n
 2550 distinct eigenvalues (or equivalently n linearly independent eigenvectors)
 2551 then the matrix is non-defective and a unique **diagonalisation/eigende-**
 2552 **composition** exists for these matrices (Theorem 4.11). In other cases we
 2553 know that a multiplicity of eigenvalues may result (see Definitions 4.13
 2554 and 4.14).

2555 Alternatively, if this square $\mathbb{R}^{n \times n}$ matrix has a non-zero determinant,
 2556 than the matrix is non-singular, i.e. an inverse matrix exists (Theorem 4.1).
 2557 Non-singular matrices are closed under addition and multiplication, have
 2558 an identity element (I) and an inverse element, thus they form a group.

2559 Note, that non-singular and non-defective matrices are not identical
 2560 sets, as for example a rotation matrix will be invertible (determinant is
 2561 non-zero) but not diagonalizable in the real numbers (non-distinct real
 2562 eigenvalues).

2563 Let us follow the branch of non-defective square $A \in \mathbb{R}^{n \times n}$ matrices.
 2564 A is normal if the condition $A^\top A = AA^\top$ holds. Moreover, if the more
 2565 restrictive condition holds $A^\top A = AA^\top = I$, then the matrix is called
 2566 orthogonal (see Definition 3.8) and is a subset of the non-singular (in-
 2567 vertible) matrices and satisfy the very useful condition $A^\top = A^{-1}$. Or-
 2568 thogonal matrices are closed under addition and multiplication, have an
 2569 identity element (I) and an inverse element, thus they also form a group.

2570 The normal matrices have a frequently encountered subset, the symmet-
 2571 ric matrices $S \in \mathbb{R}^{n \times n}$ which satisfy $S = S^\top$. Symmetric matrices have
 2572 only real eigenvalues. A subset of the symmetric matrices are the positive
 2573 definite matrices P that satisfy the condition of $x^\top Px > 0$, then a unique
 2574 a unique **Cholesky decomposition** exists (Theorem 4.17). Positive defi-
 2575 nite matrices have only positive eigenvalues and are always invertible (i.e.
 2576 have a non-zero determinant).

2577 Another subset of the symmetric matrices are the **diagonal matrices D**
 2578 in which the entries outside the main diagonal are all zero. Diagonal ma-
 2579 trices are closed under multiplication and addition, but do not necessarily
 2580 form a group (this is only the case if all diagonal entries are non-zero so
 2581 that the matrix is invertible). A prominent special case of the diagonal
 2582 matrices is the identity matrix I .

2583 4.8 Further Reading

2584 Most of the content in this chapter establishes underlying mathematics
 2585 and connects them to methods for studying mappings, many of these un-
 2586 derly machine learning at the level of underpinning software solutions and
 2587 building blocks for almost all machine learning theory. Matrix characteri-
 2588 zation using determinants, eigenspectra and eigenspaces are fundamental
 2589 features and conditions for categorizing and analyzing matrices, this ex-
 2590 tends 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 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, features vectors. Similarly, matrix decomposition methods such as Cholesky decomposition reappear often when we have to compute or simulate random events (Rubinstein and Kroese, 2016).

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, 1901a), 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,

2633 or when we want to perform data preprocessing, e.g. to decorrelate pre-
 2634 predictor variables of a design matrix (e.g. Ormoneit et al. (2001)). SVD
 2635 is the basic two-dimensional version of a more general decomposition of
 2636 data in, so called, tensors (Kolda and Bader, 2009). Tensors reflect higher-
 2637 dimensional arrays and SVD-like and low-rank approximation s on tensors
 2638 are for example the CP (Carroll and Chang, 1970) or Tucker Decomposi-
 2639 tion (Tucker, 1966).

2640 The SVD low-rank approximation is frequently used in machine learn-
 2641 ing for both computational efficiency reasons. This is because it reduces
 2642 the amount of memory and operations with non-zero multiplications we
 2643 need to perform on potentially very large matrices of data (Trefethen and
 2644 Bau III, 1997). Moreover, low-rank approximation is used to operate on
 2645 matrices that may contain missing values as well as for purposes of lossy
 2646 compression and dimensionality reduction (Moonen and De Moor, 1995;
 2647 Markovsky, 2011).

2648

Exercises

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

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \\ 0 & 2 & 4 \end{bmatrix} \quad (4.132)$$

- 4.2 Compute the following determinant efficiently.

$$\begin{bmatrix} 2 & 0 & 1 & 2 & 0 \\ 2 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -2 & 0 & 2 & -1 & 2 \\ 2 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (4.133)$$

- 2649 4.3 Let us 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

$$\mathbf{A} = \begin{bmatrix} 0 & -1 & 1 & 1 \\ -1 & 1 & -2 & 3 \\ 2 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 \end{bmatrix} \quad (4.134)$$

- 2650 4.5 Diagonalizability of a matrix is unrelated to its invertibility. Determine for
 2651 the following for matrices it if is diagonalizable and/or invertible $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,

2652 $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

- 4.6 Find the SVD of the following matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \quad (4.135)$$

4.7 Let us find the singular value decomposition of

$$\mathbf{A} = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}. \quad (4.136)$$

4.8 Find the best rank-1 approximation of

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \quad (4.137)$$