Lecture Notes
Institute for Machine Learning, Johannes Kepler University Linz

JOHANNES KEPLER UNIVERSITY LINZ

Basic Methods of Data Analysis

Winter Semester 2021/22

by Philipp Renz

May 10, 2024

Institute for Machine Learning
Johannes Kepler University Linz

Chapter 1

Linear algebra: Basics

In data analysis we will often represent our data as matrices. Consequently many algorithms are described using the concept of matrices and their properties which could be broadly called the study of linear algebra. Here I don't want to give an axiomatic introduction but rather provide some basic results with the prerequisite that the reader already knows some basics about vectors and matrices. We will denote vectors as bold lowercase letters

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \tag{1.1}$$

where a_i is the ith entry of the vector $\boldsymbol{a}$.

Matrices are denoted by bold uppercase letters

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}. \tag{1.2}$$

Depending on the context it is useful to view a matrix as a concetation of either row or column vectors.

$$\boldsymbol{A} = \begin{pmatrix} — & \boldsymbol{a}_1^r & — \\ & \vdots & \\ — & \boldsymbol{a}_m^r & — \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{a}_1^c & \dots & \boldsymbol{a}_n^c \\ | & & | \end{pmatrix}. \tag{1.3}$$

We made an explicit distinction between row- and column vectors using superscripts here but these will often be left out for better readability.

1.1 Matrix multiplication

Given two matrices

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in R^{m \times n}, \qquad \boldsymbol{B} = \begin{pmatrix} b_{11} & \dots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{np} \end{pmatrix} \in R^{n \times p} \tag{1.4}$$

their product $\boldsymbol{C} = \boldsymbol{AB}$ is of shape $(m \times p)$ and it's entries are

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} \tag{1.5}$$

The resulting matrix $\boldsymbol{C}$ has a shape of $(m \times p) \leftarrow (m \times n)(n \times p)$. The inner dimension n is "eliminated" by summing over it. One can also see that the sum moves horizontally/rowwise over the entries of A and down the columns of B

Often we are interested in a special case in which one of the matrices reduces to a vector, that means it has only one column or one row. We define a column vector as a shorthand notation for a matrix $\boldsymbol{B} \in R^{n \times 1}$ with only one column and drop the column index:

$$\boldsymbol{B} = \begin{pmatrix} b_{11} \\ \vdots \\ b_{n1} \end{pmatrix} \hat{=} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \boldsymbol{b} \tag{1.6}$$

Implicitly when we talk of a vector b with n elements we mean a matrix $\boldsymbol{B} \in R^{n \times 1}$. When we want to talk about a matrix with only one row we write this a transposed vector

$$\boldsymbol{b^T} = (b_1 \dots b_n) \hat{=} (b_{11} \dots b_{1n}) \tag{1.7}$$

1.1.1 Vector-vector products

Two vectors can be multiplied with each other in two different ways. The scalar or inner product is defined as

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \tag{1.8}$$

This can be seen as a multiplication of matrices with shapes $(1 \times m)(m \times 1) \rightarrow (1 \times 1)$ and results in a scalar value.

The outer product between two vectors is

$$\boldsymbol{C} = \boldsymbol{a}\boldsymbol{b}^T = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \begin{pmatrix} b_1 & \dots & b_n \end{pmatrix} = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \dots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \dots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_m b_1 & a_m b_2 & \dots & a_m b_n \end{pmatrix} \tag{1.9}$$

This means that every column is a multiple of $\boldsymbol{a}$ and every row a multiple of $\boldsymbol{b}$. We see that the entry $c_{ij} = a_i b_j$.

1.1.2 Matrix-vector multiplication

The product of a matrix and vector $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{b}$ again is a vector. If $\boldsymbol{A} \in R^{m \times n}$ and $\boldsymbol{b} \in R^n$ their product will be of shape $(m \times 1)$, or $\boldsymbol{c} \in R^m$.

Linear combination of columns

One can view matrix-vector product, $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{b}$, as calculating a linear combination of the columns of $\boldsymbol{A}$ weighted by the entries of $\boldsymbol{b}$:

$$\boldsymbol{Ab} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}b_j \\ \sum_{j=1}^{n} a_{2j}b_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}b_j \end{pmatrix} = \begin{pmatrix} a_{11}b_1 + a_{12}b_2 + \ldots + a_{1n}b_n \\ a_{21}b_1 + a_{22}b_2 + \ldots + a_{2n}b_n \\ \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \\ a_{m1}b_1 + a_{m2}b_2 + \ldots + a_{mn}b_n \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \boldsymbol{a}_1 b_1 + \boldsymbol{a}_2 b_2 + \ldots + \boldsymbol{a}_n b_n \\ | & | & & | \end{pmatrix} \tag{1.10}$$

$$= \boldsymbol{a}_1 b_1 + \boldsymbol{a}_2 b_2 + \ldots + \boldsymbol{a}_n b_n = \sum_{i=1}^{n} \boldsymbol{a}_i b_i \tag{1.11}$$

One prominent example where this view is useful is that if the columns of $\boldsymbol{A}$ form a basis. Then a vector $\boldsymbol{c}$ can be written as $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{c}_A$, where $\boldsymbol{c}_A$ are the coordinates of $\boldsymbol{c}$ with respect to the basis $\boldsymbol{A}$. Equivalently if the columns of $\boldsymbol{B}$ form another basis then

$$\boldsymbol{c} = \boldsymbol{A}\boldsymbol{c}_A = \boldsymbol{B}\boldsymbol{c}_B. \tag{1.12}$$

From this equation one can easily get the prescription of calculating the coordinates wrt. to $\boldsymbol{A}$ given those wrt. $\boldsymbol{B}$,

$$\boldsymbol{c}_A = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{c}_B. \tag{1.13}$$

Dot product

Alternatively we can view a matrix vector product $\boldsymbol{Ab}$ as calculating the dot products of $\boldsymbol{b}$ with each row of $\boldsymbol{A}$,

$$\boldsymbol{Ab} = \begin{pmatrix} — & \boldsymbol{a}_1 & — \\ & \vdots & \\ — & \boldsymbol{a}_m & — \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{b} \\ | \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1 \cdot \boldsymbol{b} \\ \vdots \\ \boldsymbol{a}_m \cdot \boldsymbol{b} \end{pmatrix} \tag{1.14}$$

1.1.3 Matrix-matrix multiplication

Now that we've looked at some special cases of matrix multiplication we can go back to the general case $\boldsymbol{C} = \boldsymbol{AB}$.

Recalling the definition of matrix multiplication

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk} \tag{1.15}$$

we can look at the index k and see that the k-th column of $\boldsymbol{C}$ is formed by the matrix vector product of A and the k-th column vector of $\boldsymbol{B}$,

$$\boldsymbol{C} = \boldsymbol{AB} = \boldsymbol{A} \begin{pmatrix} | & & | \\ \boldsymbol{b}_1 & \ldots & \boldsymbol{b}_p \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{A}\boldsymbol{b}_1 & \ldots & \boldsymbol{A}\boldsymbol{b}_p \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{c}_1 & \ldots & \boldsymbol{c}_p \\ | & & | \end{pmatrix}. \tag{1.16}$$

Breaking matrix-matrix multiplication down to multiple matrix vector operations allows us to view it as calculating multiple linear combinations of the columns of $\boldsymbol{A}$ or calculating the all inner products

between the rows of $\boldsymbol{A}$ and the columns of $\boldsymbol{B}$,

$$\boldsymbol{C} = \begin{pmatrix} — & \boldsymbol{a}_1 & — \\ & \vdots & \\ — & \boldsymbol{a}_m & — \end{pmatrix} \begin{pmatrix} | & & | \\ \boldsymbol{b}_1 & \dots & \boldsymbol{b}_p \\ | & & | \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1 \cdot \boldsymbol{b}_1 & \boldsymbol{a}_1 \cdot \boldsymbol{b}_2 & \dots & \boldsymbol{a}_1 \cdot \boldsymbol{b}_p \\ \boldsymbol{a}_2 \cdot \boldsymbol{b}_1 & \boldsymbol{a}_2 \cdot \boldsymbol{b}_2 & \dots & \boldsymbol{a}_2 \cdot \boldsymbol{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_m \cdot \boldsymbol{b}_1 & \boldsymbol{a}_m \cdot \boldsymbol{b}_2 & \dots & \boldsymbol{a}_m \cdot \boldsymbol{b}_p \end{pmatrix}. \tag{1.17}$$

From this we can read that $c_{ik} = \boldsymbol{a}_i \boldsymbol{b}_j$.

Sum of outer products

Another view that is often useful is to write a matrix-matrix multiplication as a sum of outer products. Looking again at our basic definition

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} = a_{i1} b_{1k} + a_{i2} b_{2k} + a_{i3} b_{3k} \tag{1.18}$$

we can see that for example the term $a_{i2} b_{2k}$ is the ik-th entry of the outer-product between the second column vector of $\boldsymbol{A}$ and the second row vector of $\boldsymbol{B}$. Thus we can deduce that the matrix $\boldsymbol{C}$ is a sum of outer-products,

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{pmatrix} | & & | \\ \boldsymbol{a}_1 & \dots & \boldsymbol{a}_n \\ | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{b}_1 & — \\ & \vdots & \\ — & \boldsymbol{b}_n & — \end{pmatrix} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{b}_i^T. \tag{1.19}$$

1.2 Orthogonal matrices

An orthogonal matrix is a square matrix with the special property that its transposed is equal to its inverse. That is for $\boldsymbol{U} \in \mathbb{R}^{m \times m}$

$$\boldsymbol{U^T U} = \boldsymbol{U U^T} = \mathbf{1} \tag{1.20}$$

$$\boldsymbol{U}^T = \boldsymbol{U}^{-1} \tag{1.21}$$

Viewing this matrix-matrix product as pairwise dot-products between the columns/rows of the two matrices reveals that the both the columns as well as the rows of $\boldsymbol{U}$ form an orthonormal set of vectors:

$$\begin{pmatrix} — & \boldsymbol{u}_1 & — \\ & \vdots & \\ — & \boldsymbol{u}_m & — \end{pmatrix} \begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_m \\ | & & | \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 \cdot \boldsymbol{u}_1 & \boldsymbol{u}_1 \cdot \boldsymbol{u}_2 & \dots & \boldsymbol{u}_1 \cdot \boldsymbol{u}_m \\ \boldsymbol{u}_2 \cdot \boldsymbol{u}_1 & \boldsymbol{u}_2 \cdot \boldsymbol{u}_2 & \dots & \boldsymbol{u}_2 \cdot \boldsymbol{u}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u}_m \cdot \boldsymbol{u}_1 & \boldsymbol{u}_m \cdot \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \cdot \boldsymbol{u}_m \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{1.22}$$

The transformation can be seen as a rotation followed by reflection. This is because angles and lengths are preserved under the transformation. To see this take any two vectors $\boldsymbol{a}, \boldsymbol{b}$ and apply the transform $\boldsymbol{U}$. We first establish that multiplication of a vector with $\boldsymbol{U}$ does not change its length:

$$\|\boldsymbol{Ua}\|^2 = \boldsymbol{a}^T \boldsymbol{U}^T \boldsymbol{U} \boldsymbol{a} = \boldsymbol{a^T 1 a} = \|\boldsymbol{a}\|^2 \tag{1.23}$$

$$\cos\left(\angle(\boldsymbol{Ua}, \boldsymbol{Ub})\right) = \frac{\boldsymbol{Ua} \cdot \boldsymbol{Ub}}{\|\boldsymbol{Ua}\| \, \|\boldsymbol{Ub}\|} \tag{1.24}$$

$$= \frac{\boldsymbol{a}^T \boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{Ub}}{\|\boldsymbol{Ua}\| \, \|\boldsymbol{Ub}\|} \tag{1.25}$$

$$= \frac{\boldsymbol{a}^T \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|} = \cos\left(\angle(\boldsymbol{a}, \boldsymbol{b})\right). \tag{1.26}$$

1.3 Eigenvalues and eigenvectors

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ can be seen as a linear map that maps a vector $\boldsymbol{v} \in \mathbb{R}^n$ to a vector $\boldsymbol{Av} \in \mathbb{R}^n$. There are some vectors called *eigenvectors* (derived from the german word "eigen") which are mapped onto a multiple of themselves,

$$\boldsymbol{Av} = \lambda \boldsymbol{v}. \tag{1.27}$$

If this equation holds $\boldsymbol{v}$ is called an eigenvector and λ its associated eigenvalue. The eigenvalues are a useful characteristic of a matrix and finding eigenvalues has lots of applications ranging from coupled oscillators, heat and Schrödinger equations in physics to principal component analysis in data analysis.

We still need a way to find the eigenvalues/vectors. From (1.27) we get

$$\boldsymbol{Av} = \lambda \mathbf{1} \boldsymbol{v} \tag{1.28}$$

$$(\boldsymbol{A} - \lambda \mathbf{1}) \boldsymbol{v} = \mathbf{0} \tag{1.29}$$

This yields a linear system of equations, which only has a solution if the columns of $\boldsymbol{A} - \lambda \mathbf{1}$ are linearly dependent. Thus we can find eigenvalues by setting the determinant of this matrix to zero

$$\det(\boldsymbol{A} - \lambda \mathbf{1}) = 0 \tag{1.30}$$

This gives a polynomial of degree n, called the characteristic polynomial, which according to the fundamental theorem of calculus has n roots and can be written as

$$\det(\boldsymbol{A} - \lambda \mathbf{1}) = \prod_{i=1}^{n} (\lambda - \lambda_i), \tag{1.31}$$

where the λ_i are solutions to the equation above. The number of times a specific value occurs in this factorization is called an eigenvalues algebraic multiplicity. After calculating the eigenvalues one can insert them into (1.29) one ofter another and solve for the eigenvectors v_i. One interesting question is whether the set of all eigenvectors form a complete basis of $\mathbb{R}^n$. This is not the case if for an eigenvalue with algebraic multiplicity $n_a > 1$ the null space of $\boldsymbol{A} - \lambda_i$ has dimension less than n_a.

Example Given a matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{1.32}$$

We can calculate it's characteristic polynomial

$$\det(\boldsymbol{A} - \lambda \mathbf{1}) = (1 - \lambda)^2 \tag{1.33}$$

Thus we get an eigenvalue $\lambda = 1$ with algebraic multiplicity of two. However when inserting this eigenvalue to solve for $\boldsymbol{v}$ we get

$$(\boldsymbol{A} - \lambda \mathbf{1})\boldsymbol{v} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \Rightarrow \boldsymbol{v} = s \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{1.34}$$

1.3.1 Diagonalization of a matrix

If we find an eigenbasis, that means a basis that only consists of eigenvectors, we can write the following:

$$\boldsymbol{A} \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \lambda_1 \boldsymbol{v}_1 & \dots & \lambda_n \boldsymbol{v}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{1.35}$$

or

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda} \tag{1.36}$$

Since the columns of $\boldsymbol{V}$ form a basis they are linear independent and one can invert $\boldsymbol{V}$. Thus if there is an eigenbasis a matrix can be "diagonalized":

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} \qquad\qquad \boldsymbol{\Lambda} = \boldsymbol{V}^{-1}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} \tag{1.37}$$

This form is often practical, for example it's easy to compute

$$\boldsymbol{A}^p = \underbrace{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}\dots\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}}_{p \text{ times}} \tag{1.38}$$

$$= \boldsymbol{V}\boldsymbol{\Lambda}^p\boldsymbol{V}^{-1} = \boldsymbol{V} \begin{pmatrix} \lambda_1^p & & \\ & \ddots & \\ & & \lambda_n^p \end{pmatrix} \boldsymbol{V}^{-1}. \tag{1.39}$$

1.4 Definiteness

A matrix $\boldsymbol{A}$ is (strictly) positive-definite if

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} > 0 \quad \forall \quad \boldsymbol{v} \neq 0 \tag{1.40}$$

Similarly $\boldsymbol{A}$ is positive semi-definite if

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \geq 0 \quad \forall \quad \boldsymbol{v} \neq 0. \tag{1.41}$$

A matrix $\boldsymbol{A}$ is (strictly) negative-definite if

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} < 0 \quad \forall \quad \boldsymbol{v} \neq 0. \tag{1.42}$$

A matrix $\boldsymbol{A}$ is negative-semidefinite if

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \leq 0 \quad \forall \quad \boldsymbol{v} \neq 0. \tag{1.43}$$

Note: Don't be confused if definitions vary. Some people say positive-definite for non-strict inequalities and then specify the strict inequality using "stricly positive-definite". Some specifically say positive-semidefinite if they mean non-strict inequality. Often the meaning is apparent from the context.

1.5 Symmetric matrices

A matrix $\boldsymbol{A}$ is symmetric if $\boldsymbol{A} = \boldsymbol{A}^T$ that is if $a_{ij} = a_{ji}$. Eigenvalues of real symmetric matrices are real. Suppose λ is a possibly complex eigenvalue of symmetric matrix $\boldsymbol{A}$ and eigenvector $\boldsymbol{v} \neq 0$.

Proof:

$$\lambda \bar{\boldsymbol{v}}^T \boldsymbol{v} = \bar{\boldsymbol{v}}^T \lambda \boldsymbol{v} = \bar{\boldsymbol{v}}^T A \boldsymbol{v} = \bar{\boldsymbol{v}}^T A^T \boldsymbol{v} = \tag{1.44}$$

$$= (A\bar{\boldsymbol{v}})^T \boldsymbol{v} = (\bar{A}\bar{\boldsymbol{v}})^T \boldsymbol{v} = (\bar{\lambda}\bar{\boldsymbol{v}})^T = \bar{\lambda}\bar{\boldsymbol{v}}^T \boldsymbol{v} \tag{1.45}$$

From this we see that $\lambda = \bar{\lambda}$, where we used the fact that $\overline{\boldsymbol{A}\boldsymbol{v}} = \bar{\boldsymbol{A}}$

For real symmetric matrices eigenvectors with differing related eigenvalues are orthogonal. To see this consider the following calculation:

$$\lambda_1 \boldsymbol{v}_1^T \boldsymbol{v}_2 = \boldsymbol{v}_1^T A^T \boldsymbol{v}_2 = \boldsymbol{v}_1^T A \boldsymbol{v}_2 = \boldsymbol{v}_1^T \boldsymbol{v}_2 \lambda_2. \tag{1.46}$$

Here we see that if $\lambda_1 \neq \lambda_2$ it follows that $\boldsymbol{v}_1 \cdot \boldsymbol{v}_2 = 0$.

If an eigenvalue is associated with multiple eigenvectors the eigenvectors span a higher dimensional eigenspace. For example, if both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ have an eigenvalue of λ we get that

$$\boldsymbol{A}(a\boldsymbol{v}_1 + b\boldsymbol{v}_2) = a\boldsymbol{A}\boldsymbol{v}_1 + b\boldsymbol{A}\boldsymbol{v}_2 = a\lambda\boldsymbol{v}_1 + b\lambda\boldsymbol{v}_2 = \lambda(a\boldsymbol{v}_1 + b\boldsymbol{v}_2). \tag{1.47}$$

Thus every linear combination of the two vectors is also an eigenvector. Use gram-schmidt to get orthogonal basis for this space.

Each symmetric real matrix has an orthogonal eigenbasis. In case all eigenvalues are different from each other this follows from above. The general case is a bit more complicated but we'll give an intuition below. To do this we first need to establish that we can always find at least one non-trivial eigenvector for a matrix $\boldsymbol{A}$. This is because the characteristic polynomial will have at least one solution. Thus we can find an eigenvector $\boldsymbol{v}_1$. Then we take any vector $\boldsymbol{u}$ that is orthogonal to $\boldsymbol{v}_1$. We observe that

$$0 = \boldsymbol{u}^T \boldsymbol{v}_1 \lambda_1 = \boldsymbol{u}^T \boldsymbol{A}\boldsymbol{v}_1 = \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{v}_1 = (\boldsymbol{A}\boldsymbol{u})^T \boldsymbol{v}_1 \tag{1.48}$$

Thus we know that $\boldsymbol{A}\boldsymbol{u}$ stays orthogonal to $\boldsymbol{v_1}$. This means that we can define a new subspace $V_1^\top$ that is orthogonal to $\boldsymbol{v}_1$, i.e. it only contains vectors orthogonal to $\boldsymbol{v}_1$. Now we know that $\boldsymbol{A}$ maps all vectors $V_1^\top$ to the same space. This means we can again find an eigenvector for this mapping. We could in principle transform the linear mapping that $\boldsymbol{A}$ represents into a basis of $V_1^\top$. Then we would have the familiar matrix form again. In this way we would get a second eigenvector $\boldsymbol{v}_2$. This can be done until one has a complete set of eigenvectors. By construction they will all be orthogonal to each other.

1.6 Singular value decomposition

The singular value decomposition is a useful way to factorize a matrix. Any matrix $\boldsymbol{M} \in R^{m\times n}$ can be written as a product of three matrices

$$\underset{(m\times n)}{\boldsymbol{M}} = \underset{(m\times m)}{\boldsymbol{U}}\underset{(m\times n)}{\boldsymbol{\Sigma}}\underset{(n\times n)}{\boldsymbol{V}^T} \tag{1.49}$$

$$, \tag{1.50}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal

$$\boldsymbol{U}\boldsymbol{U}^T = \mathbf{1}_{m\times m} \qquad\qquad \boldsymbol{V}\boldsymbol{V}^T = \mathbf{1}_{n\times n}, \tag{1.51}$$

and $\boldsymbol{\Sigma}$ is diagonal.

Without loss of generality we can assume that $m \geq n$. This would only result in empty columns instead of empty rows of $\boldsymbol{\Sigma}$. The shapes look approximately like this (5×3) in the example.

$$\begin{bmatrix} \\ \\ \boldsymbol{M} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \boldsymbol{U} \\ \\ \\ \end{bmatrix}_{m\times m} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \end{bmatrix}_{m\times n} \begin{bmatrix} \\ \boldsymbol{V}^T \\ \\ \end{bmatrix}_{n\times n} \tag{1.52}$$

Proof of existence We will prove that such a representation always exists. We know that $\boldsymbol{M}^T\boldsymbol{M}$ has an orthogonal eigenbasis as it is symmetric, and non-negative eigenvalues as it is positive-definite,

$$\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{v}_i = \alpha_i\boldsymbol{v}_i, \qquad\qquad \alpha_i \geq 0, \qquad\qquad \boldsymbol{v}_i \cdot \boldsymbol{v}_j = \delta_{ij}. \tag{1.53}$$

Next we define vectors

$$\boldsymbol{u}_i = \frac{\boldsymbol{M}\boldsymbol{v}_i}{\sqrt{\alpha_i}} \tag{1.54}$$

and observe that they are also pairwise orthogonal

$$\boldsymbol{u}_i\boldsymbol{u}_j = \frac{\boldsymbol{v}_i^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{v}_i}{\sqrt{\alpha_i\alpha_j}} = \frac{\boldsymbol{v}_i^T\boldsymbol{v}_i\alpha_i}{\sqrt{\alpha_i\alpha_j}} = \frac{\delta_{ij}\alpha_i}{\sqrt{\alpha_i\alpha_j}} = \frac{\delta_{ij}\alpha_i}{\sqrt{\alpha_i\alpha_i}} = \delta_{ij} \tag{1.55}$$

We also get that the $\boldsymbol{u}_i$ are eigenvectors of $\boldsymbol{M}\boldsymbol{M}^T$,

$$(\boldsymbol{M}\boldsymbol{M}^T)\boldsymbol{u}_i = (\boldsymbol{M}\boldsymbol{M}^T)\boldsymbol{M}\boldsymbol{v}_i\frac{1}{\sqrt{\alpha_i}} = \boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M})\boldsymbol{v}_i\frac{1}{\sqrt{\alpha_i}} = \boldsymbol{M}\alpha_i\boldsymbol{v}_i\frac{1}{\sqrt{\alpha_i}} = \alpha_i\frac{\boldsymbol{M}\boldsymbol{v}_i}{\sqrt{\alpha_i}} = \alpha_i\boldsymbol{u}_i \tag{1.56}$$

Next we assume that$\boldsymbol{M}^T\boldsymbol{M}$ has l non-zero eigenvalues. We take the respective eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l$ and compute the corresponding $\boldsymbol{u}_i$ simultaneously, while using $\sigma_i = \alpha_i^{-\frac{1}{2}}$,

$$\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_l \\ | & & | \end{pmatrix} = \boldsymbol{M}\begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_l \\ | & & | \end{pmatrix}\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix} \tag{1.57}$$

$$\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l \\ | & & | \end{pmatrix} \tag{1.58}$$

$$\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l \\ | & & | \end{pmatrix} \tag{1.59}$$

$$\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l \\ | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix} \tag{1.60}$$

To get to our desired factorization and isolate $\boldsymbol{M}$ we would like to get somehow get rid of the last two matrices on the right. The l orthonormal vectors $\boldsymbol{v}_l$ form a basis of $\mathbb{R}^<$ and $\boldsymbol{V}^T\boldsymbol{V} = \mathbf{1}_{l\times l}$. However, this does not hold for the transpose, and $\boldsymbol{V}^T\boldsymbol{V} = \mathbf{1}_{n\times n}$ does not hold in general. This means and we can't just cancel it out. We know however, that

$$\boldsymbol{v}_i^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{v_i} = \boldsymbol{v}_i^T\boldsymbol{v}_i\alpha_i \tag{1.61}$$

$$\alpha_i = 0 \;\Rightarrow\; \boldsymbol{M}\boldsymbol{v}_i = 0 \tag{1.62}$$

This allows us to do the following: We add some columns/rows to the above such that such that they form a full set of n orthonormal eigenvectors, which always exist. Then we can use the knowledge that

$$\boldsymbol{M} = \boldsymbol{M} \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l & \boldsymbol{v}_{l+1} & \dots & \boldsymbol{v}_n \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \\ — & \boldsymbol{v}_l+1 & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix} \tag{1.63}$$

$$= \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{M}\boldsymbol{v}_1 & \dots & \boldsymbol{M}\boldsymbol{v}_l & \boldsymbol{M}\boldsymbol{v}_{l+1} & \dots & \boldsymbol{M}\boldsymbol{v}_n \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \\ — & \boldsymbol{v}_l+1 & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix} \tag{1.64}$$

$$= \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{M}\boldsymbol{v}_1 & \dots & \boldsymbol{M}\boldsymbol{v}_l & \mathbf{0} & \dots & \mathbf{0} \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \\ — & \boldsymbol{v}_l+1 & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix} \tag{1.65}$$

$$= \boldsymbol{M} \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l & \mathbf{0} & \dots & \mathbf{0} \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \\ — & \boldsymbol{v}_l+1 & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix} \tag{1.66}$$

$$= \boldsymbol{M} \sum_{i=1}^{l} \boldsymbol{v}_i \boldsymbol{v}_i^T + \sum_{i=l+1}^{n} \mathbf{0}\, \boldsymbol{v}_i^T = \boldsymbol{M} \sum_{i=1}^{l} \boldsymbol{v}_i \boldsymbol{v}_i^T \tag{1.67}$$

$$= \boldsymbol{M} \begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \dots & \boldsymbol{v}_l \\ | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix} = \boldsymbol{M}. \tag{1.68}$$

For this to work we needed to know that $\boldsymbol{M}\boldsymbol{v}_i = \mathbf{0}$ for $i \in \{l+1, \dots, n\}$. This is just the rhs. of (1.60) and we get that

$$\boldsymbol{M} = \underbrace{\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l \\ | & & | \end{pmatrix}}_{\boldsymbol{U}^{(1)}} \underbrace{\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{pmatrix}}_{\boldsymbol{\Sigma}^{(1)}} \underbrace{\begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix}}_{\boldsymbol{V}^{(1)T}} \tag{1.69}$$

This is the reduced singular value decomposition of matrix. However, in many cases we would like the outer matrices to be square and orthogonal. We can achieve this via padding zeros in the diagonal matrix and adding completing the rows/columns of the outer matrices to form full bases.

We first look at the term

$$\boldsymbol{D}\boldsymbol{V}^{(1)^T} = \begin{pmatrix} | & & | \\ \boldsymbol{\sigma}_1 & \dots & \boldsymbol{\sigma}_l \\ | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \end{pmatrix} = \sum_{i=1}^{l} \boldsymbol{\sigma_i} \boldsymbol{v_i}^T \tag{1.70}$$

Similar to above we can add rows to $\boldsymbol{V}^{(1)^T}$ so that it's rows form a complete set of orthonormal vectors. If we add zero columns to Σ the overall product will not change as is evident from the outer product view of matrix multiplication.

$$\boldsymbol{D}\boldsymbol{V}^{(1)^T} = \sum_{i=1}^{l} \boldsymbol{\sigma_i} \boldsymbol{v_i}^T \tag{1.71}$$

$$= \sum_{i=1}^{l} \boldsymbol{\sigma_i} \boldsymbol{v_i}^T + \sum_{i=l+1}^{n} \mathbf{0} \boldsymbol{v_i}^T \tag{1.72}$$

$$= \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{\sigma}_1 & \dots & \boldsymbol{\sigma}_l & \mathbf{0}_{l+1} & \dots & \mathbf{0}_n \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_l & — \\ — & \boldsymbol{v}_{l+1} & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix} \tag{1.73}$$

$$= \boldsymbol{\Sigma}^{(2)} \boldsymbol{V}^T \tag{1.74}$$

We have added some columns to $\boldsymbol{\Sigma}^{(2)}$ which has a shape of $(l \times n)$. $\boldsymbol{V}$ is of shape $(n \times n)$ and now contains a complete set of orthonormal vectors. Similar to the above we can add columns to $\boldsymbol{U}^{(1)}$ and compensate them via adding zero-rows to $\Sigma^{(2)}$.

$$\boldsymbol{U}^{(1)} \boldsymbol{\Sigma}^{(2)} = \begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l \\ | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{\sigma}_1 & — \\ & \vdots & \\ — & \boldsymbol{\sigma}_l & — \end{pmatrix} \tag{1.75}$$

$$= \begin{pmatrix} | & & | & | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_l & \boldsymbol{u}_{l+1} & \dots & \boldsymbol{u}_m \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} — & \boldsymbol{\sigma}_1 & — \\ & \vdots & \\ — & \boldsymbol{\sigma}_l & — \\ — & \mathbf{0}_{l+1} & — \\ & \vdots & \\ — & \mathbf{0}_m & — \end{pmatrix} \tag{1.76}$$

$$= \boldsymbol{U\Sigma} \tag{1.77}$$

This yields the SVD in its final form:

$$\boldsymbol{M} = \boldsymbol{U\Sigma V^T} = \begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_m \\ | & & | \end{pmatrix} \begin{pmatrix} \overbrace{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{matrix}}^{l\times l} & \overbrace{\begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}}^{l\times(n-l)} \\ \underbrace{\begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}}_{(m-l)\times l} & \underbrace{\begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}}_{(m-l)\times(n-l)} \end{pmatrix} \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_n & — \end{pmatrix}, \tag{1.78}$$

The blue entries were added when we augmented the $\boldsymbol{v}_i$ and the red and green zeros when we completed the $\boldsymbol{u}_i$.

1.7 Miscellaneous

1.7.1 A real matrix with complex eigenvalues

A real matrix can have complex eigenvalues

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \vec{v} = \lambda \vec{v}$$

$$\begin{pmatrix} -\lambda & 1 \\ -1 & -\lambda \end{pmatrix} \vec{v} = 0$$

$$\det \begin{pmatrix} -\lambda & 1 \\ -1 & -\lambda \end{pmatrix} = 0$$

$$\lambda^2 + 1 = 0 \quad \lambda = \pm i \quad v_1 = \begin{pmatrix} 1 \\ i \end{pmatrix}$$

$$\begin{pmatrix} -i & 1 \\ -i & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$-ix_1 + y_2 = 0$$
$$x_1 = 1$$
$$y_2 = i$$

Note:

$$\begin{pmatrix} i & 1 \\ -1 & i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$x_i + y = 0$

$x = 1$

$y = -i$

$\vec{v_2} = \begin{pmatrix} 1 \\ -i \end{pmatrix}$

Complex number: z=a+ib Complex conjugate is defined as $\bar{z} = a - ib$ $z\bar{z} = (a + ib)(a - ib) = a^2 + b^2 \geq 0$

$$\text{Complex vector } \vec{v} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{pmatrix}$$

$\bar{\vec{v}} \cdot \vec{v} \geq 0$ with equality iff $\vec{v} = \vec{0}$

Chapter 2

Summarizing univariate data

There are n observations, where n is the number of students. It is hard to make sense of this data. First idea is analyzing the data: Calculate size of a typical person by so called "central tendency".

$$X = (160, 163, 156, \ldots, 183) = (x_1, \ldots, x_n)$$

2.1 Arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + \ldots + x_n) \tag{2.1}$$

Not a robust statistic because it is heavily influenced by outliers.

Example: 100 people live in a village. Everybody is earning 100 € a month, but one person earns 1 billion € a month. The average wage is 10099 €, but the typical person is not earning that much.

2.1.1 Median

More robust measure of central tendency. The median splits the data sample in half.

$$median(x) = \begin{cases} x_{\frac{n+1}{2}} & \text{if n is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if n is even} \end{cases} \tag{2.2}$$

Example:

$$x = (100, 110, 120, 200, 300, 1000000) \tag{2.3}$$

$$median(x) = \frac{120 + 200}{2} = 160 \tag{2.4}$$

2.1.2 Quantiles

Cut points that divide sample into equally sized slices. Quantiles divide the data into four regions.

- Q_1: the first quantile, middle number between minimum and median

- Q_2: the second quantile, median
- Q_3: the third quantile, middle number between median and maximum

Similarly deciles split the data into ten and percentiles into 100 slices.

2.1.3 Boxplot

Popular use of quantiles. It gives information both about the central tendency and spread of the data.

Whisker options:

- minimum / maximum
- $l = 1.5\ (Q_3 - Q_1)$, values outside of the whiskers are called outliers and drawn individually
- e.g. can show the 5^{th} and 95^{th} percentile

2.1.4 Histogram

More complex than boxplot but better suited for multi-modal distributions, i.e. distributions with several peaks:

- Divide data range into equally sized intervals
- Count number of samples in each interval

The interval size has to be chosen suitably.

- Too small $\Rightarrow$ all counts are either one or zero.
- Too large $\Rightarrow$ no gained information.

2.1.5 Measuring spread of data

A common way to measure the spread of data is to calculate the average distance to the mean value. A simple idea idea would be to use:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \tag{2.5}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right) \tag{2.6}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) \tag{2.7}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) - \bar{x} \tag{2.8}$$

$$= \bar{x} - \bar{x} = 0 \tag{2.9}$$

2.1.6 Mean absolute deviation

$$\text{MAD}(x) = \frac{1}{n}\sum_{i=1}^{n}|(x_i - \bar{x})| \tag{2.10}$$

Exercise: Find a that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}|(x_i - a)| \tag{2.11}$$

2.1.7 Variance: Mean squared deviation

$$\sigma^2(x) = \text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.12}$$

Alternative way to write this:

$$\begin{aligned}
\text{var}(x) =& \frac{1}{n}\left(\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right) = && (2.13)\\
=& \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i + \bar{x}^2\right) = && (2.14)\\
=& \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}^2 + \bar{x}^2\right) = && (2.15)\\
=& \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - \bar{x}^2\right) = && (2.16)\\
=& \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 = && (2.17)\\
=& \bar{x^2} - \bar{x}^2 && (2.18)
\end{aligned}$$

This means the variance can be calculated using the mean squared value and the mean of the data.

Exercise: Find the value of a that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 \tag{2.19}$$

2.1.8 Standard deviation

$$\sigma(x) = \text{std}(x) = \sqrt{\text{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2.2 Summarizing Bivariate Data

When dealing with bivariate data we have two values for each sample. For example, we measure weight (x_i) and height (y_i) of animal with index i. Our dataset is then given by tuples of data:

$$z = (z_1, z_2, \ldots, z_n) = ((x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)) \tag{2.20}$$

2.2.1 Scatterplot

2.2.2 Covariance

To analyze this kind of data we can again look at the central tendencies and spreads of the single features. Additionally we can analyze how the two features they vary together.

The covariance between x and y is defined as

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = s_{yx} \tag{2.21}$$

If the covariance is positive then on average then on avarage a positive deviation from the mean in one feature leads to one in the other and vice versa. A negative value says that a positive deviation from the mean in one feature leads to a negative one in the other.

Similar to the variance we can also write the covariance as

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) && (2.22)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\frac{1}{n}\sum_{i=1}^{n} y_i + \bar{x}\bar{y} && (2.23)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} && (2.24)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} && (2.25) \end{aligned}$$

The covariance of a feature with itself recovers the variance. The variance on covariance are often combined in the so called covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{pmatrix}$$

It gives information about spread of data (s_{xx}, s_{yy} and how they vary together.

The data is often written in form of a matrix

$$\boldsymbol{Z} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ y_1 & y_2 & \ldots & y_n \end{pmatrix} \tag{2.26}$$

and by taking the row-wise means we get the vector of means

$$\overline{\boldsymbol{z}} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

We can take another look at the covariance matrix and try to write it in the terms defined above.

$$\boldsymbol{\Sigma} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n x_i^2 - \overline{x}^2 & \sum_{i=1}^n x_i y_i - \overline{x}\,\overline{y} \\ \sum_{i=1}^n y_i x_i - \overline{x}\,\overline{y} & \sum_{i=1}^n y_i y_i - \overline{y}^2 \end{pmatrix} = \tag{2.27}$$

$$= \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i x_i & \sum_{i=1}^n y_i y_i \end{pmatrix} - \begin{pmatrix} \overline{x^2} & \overline{xy} \\ \overline{xy} & \overline{y}^2 \end{pmatrix} = \tag{2.28}$$

$$= \frac{1}{n}\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \end{pmatrix} \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix} \begin{pmatrix} \overline{x} & \overline{y} \end{pmatrix} = \tag{2.29}$$

$$= \frac{1}{n}\boldsymbol{Z}\boldsymbol{Z}^T - \overline{\boldsymbol{z}}\,\overline{\boldsymbol{z}}^T \tag{2.30}$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$$

2.2.3 Pearson Correlation Coefficient

The pearson correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{2.31}$$

can be seen as a normalized covariance. It ranges from -1 to 1. To see this, we view the quantities above as vectors.

$$u_i = (x_i - \overline{x}) \tag{2.32}$$

$$v_i = (y_i - \overline{y}) \tag{2.33}$$

$$\boldsymbol{u} = \begin{pmatrix} x_i \\ \vdots \\ x_n \end{pmatrix} - \bar{x} \qquad \boldsymbol{v} = \begin{pmatrix} y_i \\ \vdots \\ y_n \end{pmatrix} - \bar{y} \tag{2.34}$$

$$\tag{2.35}$$

$$r_{xy} = \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} = cos_\alpha$$

$$r_{xy} = \begin{cases} 1 & \text{if } x_i = ay_i + b \text{ with } a > 0 \\ -1 & \text{if } x_i = -ay_i + b \text{ with } a > 0 \end{cases} \tag{2.36}$$

For proof see Cauchy-Schwarz inequality.

2.2.4 Linear regression

In linear regression our goal is to predict the value of one variable given that of another using a linear relation. Given the data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ we want to find a function that predicts the value of y_i given that of x_i. This prediction should be made by the model

$$\hat{y}_i = wx_i + b$$

where w and b are parameters that have to be adjusted to fit the data. A simple example would be to predict the weight of a dog given its height. If all the data points would lie on a straight line we could easily calculate the values of w and b that fit the data. However, if this is not possible we can try to find parameters that minimize the mean squared error:

$$L(w, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.37}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - (wx_i + b))^2 \tag{2.38}$$

To find a minimum we first set the gradient of the loss w.r.t. to the parameters to zero.

$$\nabla L(w, b) = \begin{pmatrix} \frac{\partial L(w,b)}{\partial w} \\ \frac{\partial L(w,b)}{\partial b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

This gives us a system of linear equations:

$$\frac{\partial L}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b))x_i \tag{2.39}$$

$$\frac{\partial L}{\partial b} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b)) \tag{2.40}$$

Solving for w and b:

$$\frac{1}{n}\left(\sum_{i=1}^{n} -2(y_i - (wx_i + b))\right) = 0 \tag{2.41}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - w\frac{1}{n}\sum_{i=1}^{n} x_i - b = 0 \tag{2.42}$$

$$\bar{y} - w\bar{x} = b \tag{2.43}$$

With the result for b we can solve for w:

$$\frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b)) = 0 \tag{2.44}$$

$$\frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b))x_i = 0 \tag{2.45}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{x}\bar{y} + \bar{x}\bar{y} - w(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 + \bar{x}^2) - b\frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.46}$$

$$= s_{xy} + \bar{x}\bar{y} - ws_x - w\bar{x}^2 - b\bar{x} \tag{2.47}$$
$$| \; b = \bar{y} - w\bar{x} \; | \tag{2.48}$$
$$= s_{xy} + \bar{x}\bar{y} - ws_x - w\bar{x}^2 - \bar{x}\bar{y} + w\bar{x}^2 \tag{2.49}$$
$$= s_{xy} - ws_x \tag{2.50}$$
$$= 0 \tag{2.51}$$

Thus we get

$$s_x w = s_{xy} \tag{2.52}$$

and

$$w = \begin{cases} \frac{s_{xy}}{s_x} & \text{if } s_x \neq 0 \\ \text{arbitrary} & \text{if } s_x = 0 \end{cases} \tag{2.53}$$
$$b = \bar{y} - w\bar{x} \tag{2.54}$$

Next we want to check if this corresponds to a minimum. We do so by checking the Hessian matrix is positive-definite. The first derivatives are:

$$\frac{\partial L}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b))x_i \tag{2.55}$$
$$\frac{\partial L}{\partial b} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - (wx_i + b)) \tag{2.56}$$
$$\tag{2.57}$$

Further differentiation gives:

$$\frac{\partial^2 L}{\partial w^2} = -\frac{2}{n}\sum_{i=1}^{n} -x_i^2 = 2\overline{x^2} \tag{2.58}$$
$$\frac{\partial^2 L}{\partial w \partial b} = \frac{2}{n}\sum_{i=1}^{n} x_i = 2\bar{x} \tag{2.59}$$
$$\frac{\partial^2 L}{\partial b^2} = 2 \tag{2.60}$$

In general it holds that

$$\frac{\partial}{\partial b}\left(\frac{\partial L}{\partial w}\right) = \frac{\partial}{\partial w}\left(\frac{\partial L}{\partial b}\right), \tag{2.61}$$

which lets us save some computation.

This gives us:

$$H(L) = 2\begin{pmatrix} \overline{x^2} & \bar{x} \\ \bar{x} & 1 \end{pmatrix}$$

If this matrix is strictly positive definite we have indeed obtained a local minimum.

$$\begin{pmatrix} v_1 & v_2 \end{pmatrix}\begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \tag{2.62}$$

$$v_1^2\overline{x^2} + 2v_1v_2\overline{x} + v_2^2 \geq \tag{2.63}$$

$$v_1^2\overline{x}^2 + 2v_1v_2\bar{x} + v_2^2 = \tag{2.64}$$

$$(v_1\bar{x} + v_2)^2 \geq 0 \tag{2.65}$$

where we have used

$$\overline{x^2} - \overline{x}^2 = \text{var}(x) \geq 0 \Rightarrow \overline{x^2} \geq \overline{x}^2 \tag{2.66}$$

if var$(x) > 0$ this becomes a strict inequality which means that we get a unique minimum. Recall that when we have no variance in the x_i the value for w was arbitrary.

Thus we have established a solution for the linear regression problem.

Chapter 3

Principal component analysis

Motivating example There is a park and we are given the locations of the trees in it given two numbers. It appears that they are approximately planted in a line with slight deviations. However we would like to find a single direction which tells us the most about the location of tree. In the figure it is evident that if we can get a very good estimate of where trees are, by knowing how far to go along the red line. The (blue) distance orthogonal to it is of less importance. If we encoded each point by it's red and blue distance we would find that the red ones vary a lot more than the blue ones. Thus we will define the most important direction as the one along which the data has the highest variance.

3.1 Variance maximization

Given a vector $\boldsymbol{v}$ with unit length $||\boldsymbol{v}|| = 1$ we can calculate the coordinate wrt. to it via a dot product,

$$y_i = \boldsymbol{x}_i \cdot \boldsymbol{v}, \tag{3.1}$$

yielding a scalar for each data-point. The variance of the values $\{y_1, \ldots, y_n\}$ is

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 \tag{3.2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i\boldsymbol{v})^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{v}\right)^2 = \qquad \text{def. of } y_i \tag{3.3}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} x_{ij}v_j x_{ik}v_k - \boldsymbol{v}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{v} = \qquad \text{write dot product as sum} \tag{3.4}$$

$$= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m} v_j \sum_{i=1}^{n} x_{ji}^T x_{ik} v_k - \boldsymbol{v}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{v} = \qquad \text{rearrange} \tag{3.5}$$

$$= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m} v_j [\boldsymbol{X}^T\boldsymbol{X}]_{ik} v_k - \boldsymbol{v}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{v} = \qquad \text{replace sum by m.m.} \tag{3.6}$$

$$= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m} v_j [\boldsymbol{X}^T\boldsymbol{X}]_{ik} v_k - \boldsymbol{v}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{v} = \qquad \text{replace sum by m.m.} \tag{3.7}$$

$$= \frac{1}{n}\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{v} = \qquad \text{replace sums by m.m.} \tag{3.8}$$

$$= \boldsymbol{v}^T\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right)\boldsymbol{v} = \qquad \text{rearrange} \tag{3.9}$$

$$= \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} \qquad \text{def. of covariance} \tag{3.10}$$

Since $\boldsymbol{C}$ is real and symmetric we can express it using it's eigenvalues and eigenvectors,

$$\boldsymbol{C} = \sum_{j=1}^{m} \lambda_j \boldsymbol{w}_j \boldsymbol{w}_j^T. \tag{3.11}$$

Since $(w_1, \ldots, w_m)$ form an orthonormal basis we can also write $\boldsymbol{v}$ as linear combination of them,

$$\boldsymbol{v} = \sum_{i=1}^{m} \alpha_i \boldsymbol{w}_i. \tag{3.12}$$

Thus we shift problem of finding the entries of $\boldsymbol{v}$ to finding the α_i that maximize the variance. We can write our objective as

$$\begin{aligned}
\boldsymbol{v^T C v} &= \sum_{i=1}^{m} \alpha_i \boldsymbol{w}_i^T \sum_{j=1}^{m} \boldsymbol{w}_j \boldsymbol{w}_j^T \lambda_j \sum_{k=1}^{m} \alpha_k \boldsymbol{w}_k^T && (3.13)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_i \alpha_k \lambda_j \boldsymbol{w}_i^T \boldsymbol{w}_j \boldsymbol{w}_j^T \boldsymbol{w}_k && (3.14)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_i \alpha_k \lambda_j \delta_{ij} \delta_{jk} && (3.15)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j \lambda_j \delta_{ij} \qquad \text{sum over } k && (3.16)\\
&= \sum_{i=1}^{m} \alpha_i \alpha_i \lambda_i \qquad \text{sum over } j && (3.17)\\
&= \sum_{i=1}^{m} \alpha_i^2 \lambda_i. && (3.18)
\end{aligned}$$

Since $||\boldsymbol{v}|| = 1$ we know that $\sum_{i=1}^{m} \alpha_i^2 = 1$. The above result gives us a weighted average which is maximal when $\alpha_1 = 1$ since λ_1 is (one of) the largest eigenvalue(s). Thus the direction in which the variance of our data is maximal is just the direction of the eigenvector with the largest eigenvalue.

We now want to find the second most important direction. For this to make sense we need some additional requirement for the new direction, otherwise we could get the same one again. Since we now want to find a second vector we rename $\boldsymbol{v}$ to $\boldsymbol{v_1}$ and are looking for a vector $\boldsymbol{v_2}$ that maximizes $\boldsymbol{v}_2^T \boldsymbol{C} \boldsymbol{v}_2$, with the constraint that it has to be orthognal to $\boldsymbol{v}_1$, that is $\boldsymbol{v}_1 \cdot \boldsymbol{v}_2 = 0$. If we write the second vector in terms of the eigenvectors of $\boldsymbol{C}$,

$$\boldsymbol{v}_2 = \sum_{i=1}^{m} \beta_i \boldsymbol{w}_i, \tag{3.19}$$

then

$$\boldsymbol{v}_1^T \boldsymbol{v}_2 = \sum_{i=1}^{m} \beta_i \boldsymbol{w}_i^T \boldsymbol{w}_1 = \sum_{i=1}^{m} \beta_i \delta_{1i} = \beta_1 = 0 \tag{3.20}$$

We also have that

$$\boldsymbol{v}_2^T \boldsymbol{C} \boldsymbol{v}_2 = \sum_{i=1}^{m} \lambda_i \beta_i^2, \tag{3.21}$$

and similar to the reasoning above to maximize this under the orthogonality constraint we have to set $\beta_2 = 1$ and thus $\boldsymbol{v}_2 = \boldsymbol{w}_2$.

3.2 PCA as lossy compression

Above we tried to find directions with maximal variance. Next we want to look at PCA as lossy compression. In general we could try to find a simple rule that allows us to reconstruct our data. For example if we have bivariate data with the relation $y_i = x_i^2$ we only really need to store the x_i-values for each datapoint as we can reconstruct the y-values from it. But if we allow to general rules we run into the problem of overfitting. Thus we will allow our compression method to use a linear transformations.

We now look at what happens when we leave some of the vectors out of the picture and only include l components to encode our data.

$$\boldsymbol{x_i} \approx \boldsymbol{\mu} + \sum_{j=1}^{l} \alpha_{ij} \boldsymbol{v}_j \tag{3.22}$$

We can calculate the mean squared error of this approximation over the whole dataset.

$$\begin{aligned}
\text{MSE} =& \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \boldsymbol{\mu} + \sum_{j=1}^{l} \alpha_{ij} \boldsymbol{v_j} \right)^2 && (3.23)\\
=& \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\mu} + \sum_{j=1}^{m} \alpha_{ij} \boldsymbol{v_j} - \boldsymbol{\mu} - \sum_{j=1}^{l} \alpha_{ij} \boldsymbol{v_j} \right)^2 && (3.24)\\
=& \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=l+1}^{m} \alpha_{ij} \boldsymbol{v_j} \right)^2 && (3.25)\\
=& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=l+1}^{m} \alpha_{ij}^2 && (3.26)\\
=& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=l+1}^{m} \left((\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{v}_j \right)^2 && (3.27)\\
=& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=l+1}^{m} \left((\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{v}_j \right) \left((\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{v}_j \right) && (3.28)\\
=& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=l+1}^{m} \sum_{a=1}^{m} (x_{ia} - \mu_a) v_{ja} \sum_{b=1}^{m} (x_{ib} - \mu_b) v_{jb} && (3.29)\\
=& \sum_{j=l+1}^{m} \sum_{a=1}^{m} \sum_{b=1}^{m} v_{ja} \left(\frac{1}{n} \sum_{i=1}^{n} (x_{ia} - \mu_a)(x_{ib} - \mu_b) \right) v_{jb} && (3.30)\\
=& \sum_{j=l+1}^{m} \sum_{a=1}^{m} \sum_{b=1}^{m} v_{ja} C_{ab} v_{jb} && (3.31)\\
=& \sum_{j=l+1}^{m} \boldsymbol{v}_j^T \boldsymbol{C} \boldsymbol{v}_j && (3.32)\\
=& \sum_{j=l+1}^{m} \boldsymbol{v}_j^T \lambda_j \boldsymbol{v}_j && (3.33)\\
=& \sum_{j=l+1}^{m} \boldsymbol{v}_j^T \boldsymbol{v}_j \lambda_j && (3.34)
\end{aligned}$$

$$= \sum_{j=l+1}^{m} \lambda_j \tag{3.35}$$

$$\tag{3.36}$$

This means that the mean squared error is given by the sum of the eigenvalues of eigenvectors not used in the approximation. If we want to have low reconstruction error using only a few components we should use the ones with high eigenvalues.

Chapter 4

Multiple linear regression

Similarly to simple linear regression where again want to to predict values by fitting a linear model to observed data. But instead of just predicting the value based on one variable we will use a combination of values.

So the goal is to predict the target variable y_i given known values of a sample $\boldsymbol{x_i}$
For example we might want to predict the price y_i of a house i given its

- x_{i1}: living area
- x_{i2}: garden size
- x_{i3}: age
- x_{i4}: distance to next bus stop
- x_{i5}: baseline value or dummy variable which will be set to 1.

The last (dummy) feature takes the role of the intercept b that we used in simple linear regression and makes it easier to work out the math.

We can summarize the values of these features in a vector:

$$\boldsymbol{x_i} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \\ x_{i5} = 1 \end{pmatrix}. \tag{4.1}$$

The model predictions can be written as:

$$\begin{aligned} \hat{y_i} =& w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_4 x_{i4} + w_5 x_{i5} && (4.2) \\ =& \boldsymbol{w}^T \boldsymbol{x_i}, && (4.3) \end{aligned}$$

where the entries of the vector $\boldsymbol{w}$ should be fitted to a dataset which consists of of n tuples $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_1, y_1))$.

In matrix notation we can calculate all the predictions simultaneously:

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y_1} \\ \vdots \\ \hat{y_n} \end{pmatrix} = \begin{pmatrix} — & \boldsymbol{x}_1 & — \\ & \vdots & \\ — & \boldsymbol{x}_n & — \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{w} \\ | \end{pmatrix} = \boldsymbol{X}\boldsymbol{w}, \tag{4.4}$$

where x_{ij} is j-th feature of sample with index i. $\boldsymbol{y}$ are the true values. m is number of features. The objective is again to minimize the mean squared error,

$$L(\boldsymbol{w}) = \frac{1}{n}(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 \tag{4.5}$$

$$= \frac{1}{n}(\boldsymbol{X}\boldsymbol{w} - \hat{\boldsymbol{y}})^2 \tag{4.6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i}^T\boldsymbol{w} - y_i)^2 \tag{4.7}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} x_{ij}w_j - y_i\right)^2. \tag{4.8}$$

We again set the gradient w.r.t. to the loss to zero to find a minimum in the loss:

$$\boldsymbol{\nabla} L(\boldsymbol{w}) = \begin{pmatrix} \frac{\partial L(\boldsymbol{w})}{\partial w_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{w})}{\partial w_m} \end{pmatrix} = \begin{pmatrix} \partial_1 L(\boldsymbol{w}) \\ \vdots \\ \partial_m L(\boldsymbol{w}) \end{pmatrix}$$

The derivative of the loss w.r.t. to the a-th component of $\boldsymbol{w}$ is,

$$\frac{\partial L(\boldsymbol{w})}{\partial w_a} = \partial_a L(\boldsymbol{w}) \tag{4.9}$$

$$= \partial_a \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} x_{ij}w_j - y_i\right)^2 \tag{4.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2\left(\sum_{j} x_{ij}w_j - y_i\right)\partial_a\left(\sum_{j} x_{ij}w_j - y_i\right) \tag{4.11}$$

$$= \frac{2}{n}\sum_{i}\left(\sum_{j} x_{ij}w_j - y_i\right)\left(\sum_{j} x_{ij}\partial^a w_j\right) \tag{4.12}$$

$$= \frac{2}{n}\sum_{i}\left(\sum_{j} x_{ij}w_j - y_i\right)\left(\sum_{j} x_{ij}\partial_{ja}\right) \tag{4.13}$$

$$= \frac{2}{n}\sum_{i}\left(\sum_{j} x_{ij}w_j - y_i\right)x_{ia} = 0 \tag{4.14}$$

This gives us a system of m linear equations, one for each entry of $\boldsymbol{w}$. We can also write this in matrix form,

$$\sum_{i=1}^{n} x_{ai}^T\left(\sum_{j=1}^{m} x_{ij}w_j - y_i\right) = 0 \tag{4.15}$$

$$\sum_{i=1}^{n} x_{ai}^T\sum_{j=1}^{m} x_{ij}w_j = \sum_{j=1}^{n} x_{ai}^T y_i \tag{4.16}$$

$$\sum_{i=1}^{n} x_{ai}^T[\boldsymbol{X}\boldsymbol{w}]_i = [\boldsymbol{X}^T\boldsymbol{y}]_a \tag{4.17}$$

$$[\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}]_a = [\boldsymbol{X}^T\boldsymbol{y}]_a \tag{4.18}$$
$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{4.19}$$

We would like to solve this equation for $\boldsymbol{w}$. If $\boldsymbol{X}^T\boldsymbol{X}$ is invertible the solution is

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{4.20}$$

X^TX is not invertible if there are more features than samples or if columns of $\boldsymbol{X}$ are linear dependent. We can solve the system of equations using the SVD of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{4.21}$$

Inserting this above gives:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{4.22}$$
$$\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{w} = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \tag{4.23}$$
$$\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{w} = \boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \tag{4.24}$$
$$\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{w} = \boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \tag{4.25}$$
$$\tag{4.26}$$

There are different scenarios for the solution depending on the whether there are more samples than features and whether the feature columns are linearly independent.

More samples than features and linearly independent columns: In this case $\boldsymbol{\Sigma}$ has the following form where all $\sigma_i > 0$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ — & \mathbf{0} & — \\ & \vdots & \\ — & \mathbf{0} & — \end{pmatrix}_{n\times m} \tag{4.27}$$

$$\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{pmatrix}_{m\times m} \boldsymbol{V}^T\boldsymbol{w} = \begin{pmatrix} \sigma_1 & & & | & & | \\ & \ddots & & \mathbf{0} & \dots & \mathbf{0} \\ & & \sigma_m & | & & | \end{pmatrix}_{m\times n} \boldsymbol{U}^T\boldsymbol{y} \tag{4.28}$$

$$\boldsymbol{w} = \boldsymbol{V}\begin{pmatrix} \sigma_1^{-1} & & & | & & | \\ & \ddots & & \mathbf{0} & \dots & \mathbf{0} \\ & & \sigma_m^{-1} & | & & | \end{pmatrix}_{m\times n} \boldsymbol{U}^T\boldsymbol{y} \tag{4.29}$$

More features than samples: More features than samples will result in a $\mathbf{\Sigma}$ that includes all-zero columns. In this case the equation above will have the following form.

$$\begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_l^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}_{m \times m} \boldsymbol{w}' = \begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_l & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \\ 0 & & \dots & & & 0 \\ \vdots & & & & & \vdots \\ 0 & & \dots & & & 0 \end{pmatrix}_{m \times n} \boldsymbol{U}^T \boldsymbol{y} \tag{4.30}$$

(4.31)

We can analyse this in the following way. Since $\boldsymbol{V}$ is invertible we can solve for $\boldsymbol{w}' = \boldsymbol{V}^T \boldsymbol{w}$. We see that this system of equations is non-trivial only for the upper l rows. The lower rows just give equations of the form $0 = 0$. This also has the effect that the $w`_{l+1}, \dots, w`_m$ are cancelled out of the equation and thus become arbitrary. We can write the system of equations above for only the first l components:

$$\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_l^2 \end{pmatrix} \tag{4.32}$$

This gives us the solution for $(w'_1, \dots, w'_l)$

$$\begin{pmatrix} w'_1 \\ \vdots \\ w'_l \end{pmatrix} = \begin{pmatrix} \sigma_1 & & & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_l & 0 & \dots & 0 \end{pmatrix}_{l \times n} \boldsymbol{U}^T \boldsymbol{y} \tag{4.33}$$

$$\begin{pmatrix} w'_1 \\ \vdots \\ w'_l \end{pmatrix} = \begin{pmatrix} \sigma_1^{-1} & & & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_l^{-1} & 0 & \dots & 0 \end{pmatrix}_{l \times n} \boldsymbol{U}^T \boldsymbol{y}, \tag{4.34}$$

(4.35)

while $(w'_{l+1}, \dots, w'_m)$ are arbitrary.

Since

$$\boldsymbol{w}^T \boldsymbol{w} = \boldsymbol{w}'^T \boldsymbol{V}^T \mathbf{V} \boldsymbol{w}' = \boldsymbol{w}'^T \boldsymbol{w}' = \sum_{i=1}^{m} w_i'^2 \tag{4.36}$$

we get a minimum norm solution for $\boldsymbol{w}$ if we choose $w_i = 0, i \in \{l+1, \dots, m\}$.

Chapter 5

Probability theory

5.1 Probability

Probability is a concept to quantify the uncertainty of propositions. Probabilities are ascribed to propositions which we'll denote with capital letters. For example

A It will rain tomorrow.

B The coin toss will result in heads.

We can negate a proposition and denote it as

$\overline{\text{A}}$ It will *not* rain tomorrow.

$\overline{\text{B}}$ The coin toss will *not* result in heads.

The conjuction of two propositions is their connection via "and", and we'll write it as:

AB It will rain tomorrow *and* the coin toss will result in heads. This is often also denoted by $A \cap B$.

The disjunction of two propositions is their connection via "or", and we'll write it as:

A+B It will rain tomorrow *or* the coin toss will result in heads. This is often also denoted by $A \cup B$.

The probability of an event is given by $P(A)$. The conditional probability of A given B is written as $P(A|B)$ and denotes the probability of A given that B is true.

It follows three basic rules:

1. $P(A) = 1$: expresses certainty
2. $P(A) + P(\overline{A}) = 1$
3. $P(AB) = P(A|B)P(B) = P(B|A)P(A)$

Using these rules we can derive more complex rules. For example we can derive

$$P(A + B) = P(A) + P(B) - P(AB) \tag{5.1}$$

Proof:

$$\begin{aligned} P(A + B) &= 1 - P(\overline{AB}) && (5.2)\\ &= 1 - P(\overline{A})P(\overline{B}|\overline{A}) && (5.3)\\ &= 1 - P(\overline{A})(1 - p(B|\overline{A})) && (5.4) \end{aligned}$$

$$= 1 - P(\overline{A}) + p(\overline{A})p(B|\overline{A}) \tag{5.5}$$
$$= P(A) + P(\overline{A}B) \tag{5.6}$$
$$= p(A) + P(B)p(\overline{A}|B) \tag{5.7}$$
$$= P(A) + P(B)[1 - p(A|B)] \tag{5.8}$$
$$= P(A) + P(B) - P(A|B)P(B) \tag{5.9}$$
$$= P(A) + P(B) - P(AB) \tag{5.10}$$

If A and B are mutually exclusive $P(AB) = 0$ and we get $P(A + B) = P(A) + P(B)$.

This can be extended to three propositions:

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(CA) + P(ABC) \tag{5.11}$$

Again if A, B, C are mutually independent $P(A + B + C) = P(A) + P(B) + P(C)$. This can be extended to a general rule:

If $A_1, \ldots, A_n$ are mutually exclusive then

$$P(A_1 + \ldots + A_n) = P(A_1) + \ldots + P(A_n) \tag{5.12}$$

5.1.1 The principle of indifference

Let us pose the simple question of what probability we should assign to the event A_5, that we will throw a five on a throw of a die. Intuitively this should be one sixth. We can use what we derived above to get this result. We know for certain that some number will come up and that it will only be one, thus:

$$P(A_1 + A_2 + A_3 + A_4 + A_5 + A_6) = P(A_1) + P(A_2) + P(A_3) + P(A_4) + P(A_5) + P(A_6) = 1. \tag{5.13}$$

The principle of indifference states that if we have no reason to prefer one number over the other their probabilities should be equal thus all $P(A_i) = q$ and

$$6 * q = 1 \tag{5.14}$$
$$q = \frac{1}{6} \tag{5.15}$$

While this might seem obvious I think it's nevertheless interesting to know that many things in probability theory are based on very few assumptions.

5.1.2 Bayes' Theorem

Bayes' Theorem is often referred to in the context of updating one's beliefs in the light of new evidence. It has the following form:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{5.16}$$

It can be interpreted in the following way. We have a belief that statement B is true which is quantified by $P(B)$. This is often called the *prior*. Then we gain new information and want to update this belief to $P(B|A)$, the *posterior*.

Application of Bayes Theorem We will COVID-19 antigen tests as an example here to show how Bayes' theorem can be used. I looked up the accuracy of Covid-19 antigen tests at `https://www.medizin-transparent.at/corona-antigen-schnelltest/`.

We will use the following notations:

- $+$: positive test
- $-$: negative test
- i: infected based on PCR test
- ni: not infected based on PCR test

The accuracy of such a test is often reported in terms of two numbers. Firstly we are interested in which percentage of true positive cases the test indentifies. This is called the sensitivity. In our case the test identified 58% of the true positives (i) as positive ($+$),

$$P(+|i) = 0.58.$$

This number could easily be made to 100% by just reporting everyone as positive. For that reason we also need to know the number of false positives. In our example the false positive rate is 1%.

$$P(+|ni) = 0.01$$

These results come from a trial with the following result. Let's say we test N people with both a PCR and an antigen test then we can sort them into four categories:

- TP: True positive
- FP: False positive
- TN: True negative
- FN: False negative

	i	ni
+	TP	FP
-	FN	TN

Out of these numbers we can get the probability that the antigen test recognizes an infection if there really is one,

$$P(+|i) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.17}$$

and the probability that the test is positive although there is no infection (false positive rate):

$$P(+|ni) = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.18}$$

We next want to find out how certain we should be of an infection given a positive test.

$$P(i|+) = \frac{P(+|i)P(i)}{P(+)} \tag{5.19}$$

$$P(+) = P(+|i)P(i) + P(+|ni)P(ni) \tag{5.20}$$

The thing we are still missung is, $P(i)$, the probability of having an infection before taking the test. Without this number we cannot reach $P(i|+)$ which tells us how likely we are actually infected. This needs to be estimated first in some way. It makes sense to estimate this based on the prevalence of infected persons in the whole population. Depending on this size $P(i|+)$ will be higher or lower.

At the beginning of July 2021 the active case count was around 4000. Dividing by an approximate population of 8,000,000 we get around

$$P(i) = \frac{4,000}{8,000,000} = \frac{1}{2 \cdot 1000}. \tag{5.21}$$

On December 9th, 2021 the active case count was about 115, 000 giving

$$P(i) = \frac{115}{8000} \approx \frac{14}{1000} \tag{5.22}$$

For July this gives approximately

$$p(i|+) \approx 3\%, \tag{5.23}$$

while for December we get

$$p(i|+) \approx 42\%. \tag{5.24}$$

5.1.3 Marginalization

A useful formula referred to as marginalization is

$$P(A) = \sum_{i=1}^{n} P(AB_i), \tag{5.25}$$

which holds if $\{B_1, \ldots, B_n\}$ are exhaustive and mutually exclusive. This holds because

$$P(A) = P(A) \underbrace{\sum_{i=1}^{n} P(B_i|A)}_{=1} = \sum_{i=1}^{n} P(B_i|A)P(A) = \sum_{i=1}^{n} P(AB_i) \tag{5.26}$$

Consider for example a table which records the frequencies of the color and type of bicycles:

Bike example:

	red	blue	green
city	.22	.05	.7
mountain	.05	.18	.0
race	.13	.20	.10

Consider A= red, B_1 = city, B_2 = mountain, B_3 = race Then

$$P(\text{red}) = P(\text{red}, \text{mountain}) + P(\text{red}, \text{city}) + P(\text{red}, \text{race}) \tag{5.27}$$

5.1.4 Independence

Two propositions are called independent of each other if the value of one doesn't affect the probability of the other, which means that we don't change our belief about one given the value of the other:

$$P(A) = P(A|B) \quad \text{or} \quad P(B) = P(B|A). \tag{5.28}$$

if this holds then the joint probability is a product of the marginals.

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) = P(A)P(B) \tag{5.29}$$

In the example above we have

$$P(\text{green}) = .17 \tag{5.30}$$

$$P(\text{green}|\text{mountain}) = 0 \tag{5.31}$$

$$\tag{5.32}$$

Thus the colour and type of bicycles are not independent.

In the next example the color and type are independent given that knowledge about one doesn't say anything about the other.

	red	blue
city	$1/6$	$1/6$
race	$2/6$	$2/6$

5.1.5 Expected value

If we have probabilities over numerical variables we can calculate an expected value which intuitively is the average over many samples. Take the example of many throws of a six-sided die. We record the results

$$\overline{x} = \frac{1}{N}(1 + 5 + 6 + 2 + \ldots + 4) \tag{5.33}$$

$$= \frac{1}{N}(n_1 1 + n_2 2 + \ldots + n_6 6) \tag{5.34}$$

$$= \frac{n_1}{N}1 + \frac{n_2}{N}2 + \ldots + \frac{n_6}{N}6 \tag{5.35}$$

For large N we can assume that the relative frequencies n_i/N will approximate the probabilities p_i and we can define the expected value as

$$E[x] = \sum_{i=1}^{n} p_i x_i \tag{5.36}$$

The expected value is a linear functional:

$$E[\alpha x + \beta y] = \alpha E(x) + \beta E(y) \tag{5.37}$$

$E[x]$ is also called the population mean.

We can also compute the expected value ("average") of some function of the x_i:

$$E[f(x)] = \sum_{i=1}^{n} p_i f(x_i). \tag{5.38}$$

Example: Expected value for binomial distribution $\mathcal{B}(n, p)$. The binomial distribution models the probability to get k successes out of n tries where the probability of a success is p. The pmf is

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{5.39}$$

The expected value of this distribution is:

$$\begin{aligned}
E[x] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} && (5.40)\\
&= \sum_{k=0}^{n} k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} && (5.41)\\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} && (5.42)\\
&= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!j!} p^{j} (1-p)^{n-1-j} && (5.43)\\
&= np \sum_{j=0}^{m} \frac{m!}{(m-j)!j!} p^{j} (1-p)^{m-j} && (5.44)\\
&= np \sum_{j=0}^{m} \binom{m}{j} p^{j} (1-p)^{m-j} && (5.45)\\
&= np(p+1-p)^m && (5.46)\\
&= np && (5.47)\\
E[x] &= np && (5.48)
\end{aligned}$$

5.2 Probability density function

Generalization of probability mass function (pmf) to handle continuous values. As a motivating example consider the question of how much rain there will be tomorrow: We can specify our prediction by assigning probabilities to intervals

$$P(0\text{mm} \leq x < 1\text{mm}), P(1\text{mm} \leq x < 2\text{mm}), \ldots$$

It is often more convenient to model this by a continuous function $p(x)$. The probability that the amount of rain is in the interval $[a, b]$ can be calculated by

$$P(a \leq x \leq b) = \int_a^b p(x)\, dx \tag{5.49}$$

For a pdf the expected value is calculated using an integral:

$$E(x) = \int_{\mathcal{X}} p(x) x\, dx \tag{5.50}$$

Example : Given a Gaussian pdf

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{5.51}$$

its expected value is

$$\begin{aligned}
E[x] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx && (5.52)\\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (y+\mu) e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy && (5.53)\\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} y e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy + \mu \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy && (5.54)\\
&= 0 + \mu \int_{-\infty}^{\infty} p(y)\, dy && (5.55)\\
&= \mu && (5.56)
\end{aligned}$$

One integral vanishes as

$$\begin{aligned}
\int_{-\infty}^{\infty} y e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy = \int_{-\infty}^{0} y e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy + \int_{0}^{\infty} y e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\, dy = && (5.57)\\
= -A + A = 0 && (5.58)\\
&& (5.59)
\end{aligned}$$

The population variance is:

$$\begin{aligned}
\text{var}(x) &= E((x - E(x))^2) && (5.60)\\
&= E(x^2 - 2E(x)x + E(x)^2) && (5.61)\\
&= E(x^2) - E(E(x)x^2) + E(E(x)^2) && (5.62)\\
&= E(x^2) - 2E(x)E(x) + E(x^2) && (5.63)\\
&= E(x^2) - E(x)^2 && (5.64)
\end{aligned}$$

The population covariance is:

$$Cov(x, y) = E((x - E(x)))(y - E(x)) = E(xy) - E(x)E(y) \tag{5.65}$$

The covariance between indepedent variables is zero:

$$\begin{aligned}
Cov(x, y) &= E(xy) - E(x)E(y) && (5.66)\\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y)xy - E(x)E(y) && (5.67)\\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y)xy - E(x)E(y) && (5.68)\\
&= \sum_{x\in\mathcal{X}} p(x)x \sum_{y\in\mathcal{Y}} p(y)y - E(x)E(y) = 0 && (5.69)
\end{aligned}$$

Some rules for the covariance:

$$\begin{aligned}
&z = x + y\\
&\text{var}(z) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y)\\
&\text{var}(\alpha x) = \alpha^2 \text{var}(x)
\end{aligned}$$

If $x_1, \ldots, x_n$ are independent the variance of the sum resolves to the sum of variances:

$$\text{var}(x_1 + \ldots + x_n) = \text{var}(x_1) + \ldots + \text{var}(x_n) \tag{5.70}$$
$$\tag{5.71}$$

The expected value of the sample mean equals the population mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu \tag{5.72}$$

The variance of the sample mean is

$$\text{var}(\bar{x}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{5.73}$$
$$\tag{5.74}$$

which gives a standard deviation of $\frac{\sigma}{\sqrt{n}}$. The more samples we have the more accurate our estimate of the mean will be.

We next look at the expected value of the sample variance:

$$E(S^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2\right) \tag{5.75}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[E(x_i^2) + \frac{1}{n^2}\underbrace{E\left(\sum_j\sum_k x_j x_k\right)}_{A} - \frac{2}{n}\underbrace{E\left(\sum_{j=1}^{n} x_i x_j\right)}_{B}\right] \tag{5.76}$$

$$E(x_j x_k) = E(x_j)E(x_k) \qquad \text{if } j \neq k \tag{5.77}$$
$$E(x_j x_k) = E(x_j^2) \qquad \text{if } j = k \tag{5.78}$$

$$A = E\left(\sum_j\sum_k x_j x_k\right) \tag{5.79}$$

$$= E\left(\sum_j x_j^2 + \sum_j\sum_{k\neq j} x_j x_k\right) \tag{5.80}$$

$$= \sum_{j=1}^{n} E(x_j^2) + \sum_j\sum_{k\neq j} E(x_j)E(x_k) \tag{5.81}$$

$$= n(\mu^2 + \sigma^2) + n(n-1)\mu^2 \tag{5.82}$$
$$\tag{5.83}$$

$$B = E\left(\sum_j x_i x_j\right) = E(x_i^2) + \sum_{j \neq i} E(x_i x_j) \tag{5.84}$$

$$= \sigma^2 + \mu^2 + (n-1)\mu^2 \tag{5.85}$$

$$= \sigma^2 + n\mu^2 \tag{5.86}$$

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2 + \frac{1}{n^2}(n(\mu^2+\sigma^2) + n(n-1)\mu^2 - \frac{2}{n}(\sigma^2 + n\mu^2)\right) \tag{5.87}$$

$$= \ldots = \frac{n-1}{n}\sigma^2 \tag{5.88}$$

This shows that the expected value of the sample variance is an underestimation of the population variance.

A common way to evaluate an estimator is to look at the mean squared error, which can be decomposed into a bias and a variance part.

$$\text{MSE}(\hat{\theta}) = E[(\theta - \hat{\theta})^2] \tag{5.89}$$

$$= E[\theta^2 + \hat{\theta}^2 - 2\hat{\theta}\theta] \tag{5.90}$$

$$= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 + E[\hat{\theta^2}] - 2E[\hat{\theta}]\theta \tag{5.91}$$

$$= \underbrace{\text{var}[\hat{\theta}]}_{\text{Variance}} + \underbrace{(\theta - E(\hat{\theta}))^2}_{\text{Bias}^2} \tag{5.92}$$

$$\tag{5.93}$$

The mean squared errors of the two estimators for the variance are

$$\text{MSE}(S^2_{corr}) = \frac{2\sigma^4}{n-1} \tag{5.94}$$

$$\text{MSE}(S^2) = \frac{2n-1}{n^2}\sigma^4 = \frac{2 - \frac{1}{n}}{n}\sigma^4 \tag{5.95}$$

The estimator with the minimal mean squared error is

$$S^2_{min} = \frac{1}{n+1}\sum_i (x_i - \bar{x})^2 \tag{5.96}$$

and it's expected error is

$$\text{MSE}(S^2_{min}) = \frac{2\sigma^4}{n+1} \tag{5.97}$$

We get that:

$$\text{MSE}(S^2_{\text{corr}}) > \text{MSE}(S^2) > \text{MSE}(S^2_{\text{min}}) \tag{5.98}$$

5.3 Introduction to statistical hypothesis testing

Motivating example Consider the following problem setting. You are tossing a coin and I claim that I can predict the outcome without looking at the coin. You do not believe me and we decide to make an experiment. We throw a coin 5 times and I guess correctly every time. The question is whether you should believe that I really can predict the outcomes or if I was just lucky.

In the framework of null hypothesis significance testing this is solved in the following way. We first pose two hypotheses:

- H_0: Philipp only randomly guesses $\Rightarrow$ Null hypothesis.
- H_1: Philipp can really predict the outcome $\Rightarrow$ research hypothesis.

Then we look at the probability of the data given that the null hypothesis, H_0, is true:

$$P(\text{5 correct guesses}|H_0) = \frac{1}{2}^5 = \frac{1}{32} = 0.03125 \tag{5.99}$$

The lower this probability the more unlikely it gets that I got five correct guesses just by chance. If this probability is smaller than some predefined significance level α the null hypothesis is rejected and we accept the research hypothesis.

5.3.1 Bayesian hypothesis testing

However, this reasoning is sometimes criticized. The question is to which extent we believe the two hypotheses to be true. Intuitively you wouldn't believe my outrageous claim on the basis of a streak of five correct guesses. So this prior knowledge of the hardness of the task should somehow influence our position. In Bayesian hypothesis testing we would also look at

$$P(\text{5 correct guesses}|H_1) = 1 \tag{5.100}$$

Finally we would get the probabilities for our hypotheses by using Bayes' theorem

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)} \tag{5.101}$$

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)} \tag{5.102}$$

Now the only problem left is to encode our prior belief into the probabilities $P(H_0), P(H_1)$. It is not straightforward how to best do this.

5.3.2 Problematic example of NHST

Revisit the covid testing example:

- i: infected
- ni: not infected
- $+$: positive test
- $-$: negative test

Our null hypothesis, H_0 is that we are not infected and H_1 is that we are infected. Consider the case in which we get a positive test.

$$P(+|\text{ni}) = P(\text{data}|H_0) = \text{FPR} = 1\% \tag{5.103}$$

Thus we can reject the null hypothesis at a significance level of 1%. That is that given that the null hypothesis is true the probability of the data is less or equal than 1%.

However, the quantities we are actually interested in are,

$$P(\text{ni}|+), P(\text{ni}|-), P(\text{i}|+), P(\text{i}|-) \tag{5.104}$$

To get to these probabilities we need a reasonable prior that is values for $P(\text{ni})$ and $P(\text{i})$. A reasonable value for $P(\text{i})$ is the prevalence of the disease in the population. we needed the prevalence. With the data for July from above we get

$$P(i|+) = 3\% \tag{5.105}$$

Thus we would reject the null-hypothesis of not being infected with high a significance although the probability of being infected is relatively low.