# Contents

1 First steps in exploratory data analysis 1
    1.1 Distribution of single variables .............................. 1
        1.1.1 Numerical summaries .................................. 1
        1.1.2 Graphs ................................................. 6
    1.2 Joint distribution of two variables ........................... 10
        1.2.1 Numerical summaries .................................. 10
        1.2.2 Graphs ................................................. 14
    1.3 Simple preprocessing ........................................... 15
        1.3.1 Simple outlier detection ............................... 15
        1.3.2 Data standardisation .................................. 15

2 Principal component analysis 19

3 Dimensionality reduction 33

A Linear algebra 45
    A.1 Matrices ................................................. 45
    A.2 Vectors ................................................. 46
    A.3 Matrix operations as operations on column vectors ............ 47
    A.4 Orthogonal basis ......................................... 49
    A.5 Subspaces ................................................. 50
    A.6 Orthogonal projections .................................... 51
    A.7 Singular value decomposition ................................ 51
    A.8 Eigenvalue decomposition .................................. 52
    A.9 Positive semi-definite and definite matrices ................. 53
    A.10 Matrix approximations .................................... 53
Chapter 1
First steps in exploratory data analysis

This chapter is about the first steps of exploratory data analysis. It is assumed that we have available $n$ data points $x_1, \ldots, x_n$ each containing $d$ attributes and that the data have been transformed to numbers. Each data point $x_i$ thus is a $d$ dimensional vector. We first explain ways to describe any of the $d$ attributes in isolation and then methods to describe the joint behaviour of two attributes at a time. This is followed by some elements of preprocessing for further analysis of the data.

1.1 Distribution of single variables

We are here concerned with the properties of single attributes, e.g. the 10-th coordinate of the multivariate data $x_1, \ldots, x_n$. By extracting the coordinate of interest from each data point $x_i$, we obtain the univariate data $x_1, \ldots, x_n$ that we will describe using graphs or numerical summaries.

1.1.1 Numerical summaries

Summaries that characterise the typical value (location), the variability (scale), as well as symmetry and tail-behaviour (shape) of the data are presented.

Location

The perhaps simplest summary is the average value $m$ of the data

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i. \quad (1.1)$$

It is a measure of the typical or central value of the data. We also say that it measures the “location” of the data. The average value is further called the sample mean.

Assuming that the data were drawn from a random variable $x$ with probability density function $p(.)$, the average value $m$ of the data is an estimate of the mean
2 First steps in exploratory data analysis

or expected value of \( x \),

\[ E(x) = \int \xi p(\xi) d\xi. \] (1.2)

There are other ways to estimate \( E(x) \) or to measure the location of the data. The sample median is defined to be the centre of the data: equal proportions of the \( x_i \) lie above and below the median. It can be computed by ordering the \( x_i \) from small to large and then taking the middle value. Let \( x_{(i)} \) denote the ordered data, i.e.

\[ x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}, \] (1.3)

the median is then given by

\[
\text{median}(x_1, \ldots, x_n) = \text{median}(x_i) = \begin{cases} 
 x_{((n+1)/2)} & \text{if } n \text{ is odd} \\
 \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even}
\end{cases}
\] (1.4)

In contrast to the sample mean, the sample median is robust to outliers. Assume that a data point is recorded as \( x_i + \delta \) rather than as \( x_i \) because of a malfunction of the measurement device. The average then changes from \( m \) to \( m + \delta/n \), which can be arbitrarily large, while the sample median changes at most to a neighbouring data point. Let, for example, the ordered data be

\[
(1, 1, 3, 3, 5, 7, 9, 11, 18, 20)
\] (1.5)

The sample average is 7.8 while the median is \((5 + 7)/2 = 6\). If the data point 20 gets wrongly recorded as 200, the mean changes to 25.8 while the median stays the same. If 7 changes to 7000, the sample mean is 707.1 while the median changes only from 6 to 7.

Another measure of the location of the data that is more robust than the average is the trimmed average. It is the average of the data when leaving out the smallest and largest \( k < n/2 \) values,

\[
\frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}.
\] (1.6)

If \( k = 0 \), the trimmed average is the usual sample average. As \( k \) approaches \( n/2 \), the trimmed average approaches the median.

The sample average (mean) has the advantage that it can be easily computed in parallel for all dimensions of the multivariate data \( x_i \). Let \( m_1, \ldots, m_d \) be the average values of the different dimensions of \( x_i \) computed as in (1.1). We then have

\[
m = \begin{pmatrix} m_1 \\
m_2 \\
\vdots \\
m_d \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} x_i.
\] (1.7)

The sample mean \( m \) can be written as a matrix-vector multiplication. While computationally not an efficient way to compute \( m \), it is helpful in analytical work. Let \( X \) be the \( d \times n \) data matrix with the \( x_i \) in its columns,

\[
X = (x_1, \ldots, x_n)
\] (1.8)
We can then write \( m \) as

\[
m = \frac{1}{n} \mathbf{1}_n,
\]

where \( \mathbf{1}_n \) is a column vector containing \( n \) ones, i.e. \( \mathbf{1}^\top = (1, \cdots , 1)^\top \) (see e.g. Section A.3).

**Scale**

A basic way to measure the scale of the data is to determine how much, on average, they deviate from the average value,

\[
v = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - m^2.
\]

This measure is known as the sample variance. It is an estimator of the variance of the random variable \( x \),

\[
\mathbb{V}(x) = \int (\xi - \mathbb{E}(x))^2 p(\xi) d\xi = \mathbb{E}(x^2) - \mathbb{E}(x)^2.
\]

The variance is the 2nd central moment of \( x \) – the mean is the first moment while \( \mathbb{E}(x^2) \) is the second moment. Other estimators of the variance divide by \( n - 1 \) rather than \( n \). Note that both the variance and the sample variance have different units than the data. Hence, it is often better to indicate their square root. The square root of the (sample) variance is called the (sample) standard deviation.

Because of the squaring, the sample variance is more affected by outliers than the sample mean. The median can be used to obtain a more robust measure of the scale of the data: instead of measuring the average squared deviation from the average, we measure the median absolute deviation from the median (MAD),

\[
\text{MAD} = \text{median} \left( |x_i - \text{median}(x_i)| \right)
\]

This measure has the same units as the \( x_i \) themselves.

The range \( x_{(n)} - x_{(1)} \), i.e. the difference between the largest and the smallest value, is another measure of the scale the data, but it is not robust. A more robust quantity is the difference between the upper and lower end of what makes the central 50% of the data. This is called the interquartile range (IQR). Half of the data are between \( x_{([n/4])} \) and \( x_{([3/4n])} \) so that

\[
\text{IQR} = x_{([3/4n])} - x_{([n/4])},
\]

where \([3/4n]\) means \(3/4n\) rounded up to the next higher integer. The number \( x_{([1/4n])} \) is called the first quartile and often denoted by \( Q_1 \), and \( x_{([3/4n])} = Q_3 \) is the third quartile. The second quartile \( Q_2 \) is the median. For example,

\[
\begin{pmatrix}
1 & 1 & 3 & 3 & 5 & 7 & 9 & 11 & 18 & 20
\end{pmatrix}
\]

so that IQR = 8, while the sample standard deviation is \( \sqrt{v} = 6.76 \) and MAD = 4.
The median and the quartiles are examples of sample quantiles. The \( q_\alpha \)-th sample quantile is roughly the data point with a proportion \( \alpha \) of the ordered data \( x_i \) to its left, i.e. \( q_\alpha \approx x_{\lfloor n\alpha \rfloor} \). For example, the minimum and the maximum are the 0 and 1 quantiles \( q_0 \) and \( q_1 \); the median the 0.5 quantile \( q_{0.5} \) and the quartiles \( Q_1 \) and \( Q_3 \) are equal to \( q_{0.25} \) and \( q_{0.75} \) respectively. Like for the median, quantiles are often computed by interpolation if \( \alpha n \) is not an integer.

**Shape**

Skewness is a quantity that measures the symmetry of the data. For a random variable \( x \), skewness is defined as its third standardised moment,

\[
skew(x) = \mathbb{E} \left[ \left( \frac{x - \mathbb{E}(x)}{\sqrt{\mathbb{V}(x)}} \right)^3 \right] = \mathbb{E} \left[ \left( \frac{x - \mu}{\sigma} \right)^3 \right],
\]

with mean \( \mu = \mathbb{E}(x) \) and standard deviation \( \sigma = \sqrt{\mathbb{V}(x)} \). The subtraction of the mean and the division by the standard deviation normalises \( x \) to have zero mean and unit variance, which “standardises” \( x \) by removing the location and scale information before taking the third power. Removal of the scale information eases the comparison of skewness between different random variables.

A random variable that is symmetric around its mean has skewness zero. As

\[
skew(x) = \int \left( \frac{\xi - \mu}{\sigma} \right)^3 p(\xi) d\xi
\]

\[
= \int_{\xi \leq \mu} \left( \frac{\xi - \mu}{\sigma} \right)^3 p(\xi) d\xi + \int_{\xi > \mu} \left( \frac{\xi - \mu}{\sigma} \right)^3 p(\xi) d\xi
\]

skewness is positive if \( x \) tends to take on values much larger than the mean (heavy upper tails). Conversely, skewness is negative in case of heavy lower tails. For a data set \( x_1, \ldots, x_n \), skewness can be measured by replacing the expectations with sample averages, as we have done in case of the mean and variance. The measure is then called sample skewness.

Due to the third power, sample skewness is sensitive to outliers. A more robust measure can be obtained by means of the quartiles,

Galton’s measure of skewness \( = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \). (1.18)

The denominator is the interquartile range and normalises the skewness measure like the standard deviation in (1.15). By definition of the quartiles both \( Q_3 - Q_2 \) and \( Q_2 - Q_1 \) are positive. The first term measures the range of the third quarter while the second term measures the range of the second quarter. Galton’s skewness thus computes the difference between the ranges of the two quarters in a normalised way. It is positive if the range of the third quarter is larger than the range of the first quarter, and conversely. Figure 1.1 shows an example.

The fourth standardised moment is called the kurtosis,

\[
kurt(x) = \mathbb{E} \left[ \left( \frac{x - \mu}{\sigma} \right)^4 \right].
\]

(1.19)
Due to the fourth power, kurtosis is insensitive to the symmetry of the distribution of $x$. It measures how often $x$ takes on values that are considerably larger or smaller than its standard deviation; it is said to measure how heavy the tails of the distribution of $x$ are. Figure 1.2 shows the function $u \mapsto u^4$. It is relatively flat for $-1 < u < 1$ so that kurtosis basically ignores the behaviour of $x$ within one standard deviation around its mean. The function then grows rapidly and values larger than two standard deviations away from the mean contribute strongly to the value of kurtosis.

A Gaussian random variable with probability density function

$$p(\xi) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(\xi - \mu)^2}{2\sigma^2}\right)$$

has kurtosis equal to three. This is often taken as a reference value, and people report the excess kurtosis,

$$\text{excess kurtosis}(x) = \text{kurt}(x) - 3.$$  \hspace{1cm} (1.21)

Positive excess kurtosis indicates that the random variable has heavier tails than a Gaussian random variable. This means that it produces large values more often than a Gaussian. Negative excess kurtosis, on the other hand, indicates a random variable with fewer and less extreme values than a Gaussian. Figure 1.3 shows examples. Both (excess) kurtosis and skewness can thus be taken as measures of non-Gaussianity of the data.

Kurtosis can be computed by replacing the expectation with a sample average, or more robustly, by means of quantiles,

$$\text{robust kurtosis}(x) = \frac{(q_{7/8} - q_{5/8}) + (q_{3/8} - q_{1/8})}{q_{3/4} - q_{1/4}}.$$  \hspace{1cm} (1.22)
6 First steps in exploratory data analysis

Figure 1.2: The figure shows the function \( u \mapsto u^4 \) that occurs in the definition of the kurtosis. It is relatively flat on \((-1, 1)\) and grows quickly for values outside the interval.

Figure 1.3: Probability density functions (pdfs) with zero, negative, and positive excess kurtosis. The curves shown are correspondingly the Gaussian (blue), uniform (green), and Laplace (red) probability density functions.

The denominator is the interquartile range, and the numerator measures the length of the upper and lower tails. Further robust measures of kurtosis and skewness are discussed by Kim and White (2004).

1.1.2 Graphs

Graphs often allow us to see interesting patterns in the data more easily than numerical summaries.

Histogram

The histogram is among the most used graphs. To construct the histogram, we count how many times certain values occur among the \( x_i \). If the \( x_i \) can only take on finitely many values \( v_1, \ldots, v_k \), the histogram plots the frequency of occurrence
1.1 Distribution of single variables

\[ f_1, \ldots, f_k \] of each value, i.e.

\[ f_j = \frac{n_j}{n}, \quad n_j = \sum_{i=1}^{n} 1_{\{v_j\}}(x_i), \quad 1_{\{v_j\}}(\xi) = \begin{cases} 1 & \text{if } \xi = v_j \\ 0 & \text{otherwise} \end{cases} \quad (1.23) \]

The sum \( \sum_{i=1}^{n} 1_{\{v_j\}}(x_i) \) equals the number of times the value \( v_j \) occurs in the data set. Here, the values \( v_j \) do not need to be numerical.

If the \( x_i \) are real numbers, we can quantise them so that they take on only \( k \) different values, and then proceed as above. The usual way to do this is via binning: We form \( k \) bins \( B_1, \ldots, B_k \) and count the number of data points \( n_j \) that fall into each bin \( B_j \),

\[ n_j = \sum_{i=1}^{n} 1_{B_j}(x_i), \quad 1_{B_j}(\xi) = \begin{cases} 1 & \text{if } \xi \in B_j \\ 0 & \text{otherwise} \end{cases} \quad (1.24) \]

We can then compute the frequencies \( f_j = n_j/n \) as before. The bins are often chosen to have equal width \( h \). For equal bin sizes the bins \( B_j \) are

\[ B_1 = [L, L+h), \quad B_2 = [L+h, L+2h), \ldots, \quad B_k = [L+(k-1)h, L+kh]. \quad (1.25) \]

They are centred at the locations \( L + h/2 + jh \). The starting value \( L \), the bin-width \( h \) and the number of bins \( k \) are parameters that the user needs to choose. One may choose \( L \) to correspond to the smallest value of all \( x_i \) and given \( k \), the bin-width \( h \) such that the bins cover the whole range of the data.

Figure 1.4(a) shows an example. We can see that different starting values \( L \) may lead to differently looking histograms.

The frequencies can further be converted to probability density estimates by dividing them by the length of the bins. The estimated probability density function \( \hat{p}(x) \) is piecewise constant.\(^1\) For equal bin sizes \( h \), we have

\[ \hat{p}(x) = \frac{f_j}{h} = \frac{n_j}{nh} \quad \text{if } x \in B_j. \quad (1.26) \]

This is really just a re-normalisation of the frequencies, but the interpretation is a bit different. We can think that we first discretise \( x \) to equal the bin centre of \( B_j \), then count the number of data points \( n_j \) in its neighbourhood of size \( h \), and finally normalise the counts by \( nh \) to convert them to a density.

**Kernel density estimate**

A critique of density estimates like (1.26) is that they depend on the starting point \( L \), and that they are not smooth even if the data were drawn from a distribution with a smooth probability density function. Kernel density estimation addresses these two shortcomings of histograms. It still requires the specification of \( h \), which is called the bandwidth in kernel density estimation.

\(^1\)We have here overloaded the symbol \( x \) to refer to both the random variable and the value it may take. This is done very often, and also in this lecture without any further warning.
The dependency on the starting point $L$ can be removed by not discretising $x$ in (1.26). We just directly count the number of points in the neighbourhood of $x$, which can be done by computing

$$n(x) = \sum_{i=1}^{n} \nabla_h(x - x_i), \quad \nabla_h(\xi) = \begin{cases} 1 & \text{if } \xi \in [-\frac{h}{2}, \frac{h}{2}] \\ 0 & \text{otherwise.} \end{cases}$$

(1.27)

The function $\nabla_h(.)$ is called the rectangular or the boxcar function. Normalising the counts as before yields

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \nabla_h(x - x_i).$$

(1.28)

This quantity is a kernel density estimate, and the function $1/h \nabla_h(.)$ is called the boxcar kernel.

When we use the boxcar kernel, we stop counting data points as soon as they are further than a $h/2$ distance away from $x$. In a sense, the data points are assigned zero weight if outside the neighbourhood and unit weight when inside. Instead of this binary weight assignment, it is often more reasonable to assign to each data point a weight that decreases smoothly with the distance from the query point $x$. This can be done by replacing the boxcar kernel with another kernel $K_h(.)$ so that the kernel density estimate is

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i).$$

(1.29)

The kernel $K_h(.)$ must integrate to one for $\hat{p}(x)$ to integrate to one. A popular choice for $K_h(.)$ is the Gaussian kernel

$$K_h(\xi) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{\xi^2}{2h^2}\right).$$

(1.30)

Other kernels are plotted in Figure 1.5. Figure 1.6 shows the densities estimated by rescaling the histogram, with the boxcar kernel, and with the Gaussian kernel.
1.1 Distribution of single variables

Figure 1.5: Kernels used in kernel density estimation. Figure from https://en.wikipedia.org/wiki/Kernel_(statistics).

Figure 1.6: Estimating the probability density function from (a) the histogram or (b-c) via kernel density estimation.
First steps in exploratory data analysis

Figure 1.7: Boxplot and comparison to the Gaussian probability density function. Figure from [https://en.wikipedia.org/wiki/Box_plot](https://en.wikipedia.org/wiki/Box_plot)

Boxplot

The boxplot visualises several key quantile-based summaries in a single graph. Figure 1.7 shows the boxplot (top) and maps the summaries to the corresponding areas of the Gaussian distribution. Violin plots are a modification that combine quantile information with a kernel density estimate (Hintze and Nelson, 1998).

1.2 Joint distribution of two variables

The previous section described the behaviour of single variables in isolation. It was concerned with summarising and estimating the probability distribution of a single attribute of the multivariate data. This section is about numerical summaries and graphs for describing the joint probability distribution of two variables.

1.2.1 Numerical summaries

We present measures that quantify the tendency of two variables to vary together.
**Linear relationship: covariance and correlation**

Let $x$ and $y$ be two random variables with mean $\mu_x$ and $\mu_y$, respectively. The covariance $\text{cov}(x, y)$ measures the strength of association between them. It is defined as

$$\text{cov}(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)] = \mathbb{E}(xy) - \mu_x \mu_y.$$  \hfill (1.31)

From the definition, we have $\text{cov}(x, x) = \mathbb{V}(x)$ and $\text{cov}(ax + b, y) = a \text{cov}(x, y)$. The value of the covariance thus depends on the scale of $x$ and $y$.

A measure of association that does not depend on their scale is the correlation coefficient $\rho$,

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\mathbb{V}(x) \mathbb{V}(y)}}.$$  \hfill (1.32)

Denoting the standard deviation of $x$ and $y$ by $\sigma_x$ and $\sigma_y$, the correlation coefficient can be written as

$$\rho(x, y) = \mathbb{E} \left[ \left( \frac{x - \mu_x}{\sigma_x} \right) \left( \frac{y - \mu_y}{\sigma_y} \right) \right],$$  \hfill (1.33)

which means that $x$ and $y$ are standardised to zero mean and unit variance before the expectation is taken. If $\rho$ is positive, $x > \mu_x$ and $y > \mu_y$ tend to co-occur, and $x$ and $y$ are said to be positively correlated. If $\rho < 0$, $x$ and $y$ are negatively correlated. If $\rho = 0$, the random variables are uncorrelated, and $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$.

Assume $y = ax + b$, where $a \neq 0$ and $b$ are some constants. Then $\mu_y = a \mu_x + b$ and the variance of $y$ is

$$\mathbb{V}(y) = \mathbb{E}((y - \mu_y)^2) = \mathbb{E}((ax + b - a \mu_x - b)^2) = \mathbb{E}(a^2(x - \mu_x)^2) = a^2 \mathbb{V}(x).$$  \hfill (1.34)

The covariance is

$$\text{cov}(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)] = \mathbb{E}[(x - \mu_x)(ax + b - a \mu_x - b)] = a \mathbb{V}(x)$$  \hfill (1.35)

so that

$$\rho(x, y) = \frac{a \mathbb{V}(x)}{\sqrt{\mathbb{V}(x) a^2 \mathbb{V}(x)}} = \frac{a}{\sqrt{a^2}} = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$  \hfill (1.36)

That is, if $y$ is linearly related to $x$, the correlation coefficient is one for positive slopes $a$, and minus one for negative slopes. The magnitude of $a$ does not affect the value of $\rho$.

What does affect the value of $\rho$? Let $y = x + n$ where $x$ and $n$ are uncorrelated. We then have $\mathbb{V}(y) = \mathbb{V}(x) + \mathbb{V}(n)$, which we denote by $\sigma^2_x + \sigma^2_n$, and $\text{cov}(x, y) = \mathbb{V}(x) = \sigma^2_x$, so that

$$\rho(x, y) = \frac{\sigma^2_x}{\sqrt{\sigma^2_x (\sigma^2_x + \sigma^2_n)}} = \frac{1}{\sqrt{1 + \frac{\sigma^2_n}{\sigma^2_x}}}.$$  \hfill (1.37)

Hence, $|\rho|$ becomes small if $\sigma^2_n$ dominates $\sigma^2_x$, that is, if the so-called signal-to-noise ratio $\sigma^2_x / \sigma^2_n$ is small. Importantly, $\rho \approx 0$ only means that $x$ and $y$ are not linearly related. It does not mean that $x$ and $y$ are not related, see Figure 1.8.

The covariance and correlation coefficient can be estimated from $n$ tuples $(x_i, y_i)$, that is from data $(x_1, y_1), \ldots, (x_n, y_n)$, by replacing the expectation with the sample average as we have done before.

---

*Data Mining and Exploration, Spring 2017*
12 First steps in exploratory data analysis

Figure 1.8: Correlation coefficients for different data sets. Figure from https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

Covariance matrix

For a random vector \( \mathbf{x} = (x_1, \ldots, x_d)^\top \), the covariance matrix is the \( d \times d \) matrix \( \mathbf{C} = \mathbb{V}(\mathbf{x}) \) that contains all covariances of the variables,

\[
\mathbf{C} = \begin{pmatrix}
\text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_d) \\
\text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \cdots & \text{cov}(x_2, x_d) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(x_d, x_1) & \text{cov}(x_d, x_2) & \cdots & \text{cov}(x_d, x_d)
\end{pmatrix}
\] (1.38)

Since \( \text{cov}(x_i, x_j) = \text{cov}(x_j, x_i) \), \( \mathbf{C} \) is symmetric, and since \( \text{cov}(x_i, x_i) = \mathbb{V}(x_i) \), the elements on the diagonal are the variances.

Let \( \mathbf{\mu} \) be the \( d \)-dimensional vector with the means of the random variables \( x_i \). We can then write \( \mathbf{C} \) as

\[
\mathbf{C} = \mathbb{E} \left( (\mathbf{x} - \mathbf{\mu})(\mathbf{x} - \mathbf{\mu})^\top \right). 
\] (1.39)

This follows immediately from the properties of the outer product \( \mathbf{a} \mathbf{b}^\top \) between two vectors \( \mathbf{a} \) and \( \mathbf{b} \), see e.g. Section A.2. Indeed, \( (\mathbf{x} - \mathbf{\mu})(\mathbf{x} - \mathbf{\mu})^\top \) is a \( d \times d \) matrix where the \( (i, j) \)-th element is \( (x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j)) \). Its expectation is hence \( \text{cov}(x_i, x_j) \). By linearity of expectation

\[
\mathbf{w}^\top \mathbf{C} \mathbf{w} = \mathbb{E} \left( \underbrace{\mathbf{w}^\top (\mathbf{x} - \mathbf{\mu})}_{\text{scalar}} \underbrace{(\mathbf{x} - \mathbf{\mu})^\top}_{\text{scalar}} \mathbf{w} \right) = \mathbb{E} \left( (\mathbf{w}^\top (\mathbf{x} - \mathbf{\mu}))^2 \right) \geq 0,
\] (1.40)

which means that \( \mathbf{C} \) is a positive semi-definite matrix. It thus has an eigenvalue decomposition \( \mathbf{C} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top \), where \( \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d) \) is a diagonal matrix containing the eigenvalues \( \lambda_i \geq 0 \), and \( \mathbf{U} \) is an orthogonal matrix with the eigenvectors as columns (see e.g. Appendix A for a linear algebra refresher). The total variance of the \( d \) random variables \( x_i \) is the sum of all eigenvalues. With
the definition of the trace a matrix, see e.g. (A.6), we have
\[ \sum_{i=1}^{d} \mathbb{V}(x_i) = \text{trace}(C) \] (1.41)
\[ = \text{trace}(U \Lambda U^\top) \] (1.42)
\[ \overset{(A.8)}{=} \text{trace}(\Lambda U^\top U) \] (1.43)
\[ = \text{trace}(\Lambda) \] (1.44)
\[ = \sum_{i=1}^{d} \lambda_i, \] (1.45)
where we have used that, for an orthogonal matrix \( U \), \( U^\top U \) is the identity matrix.

If \( x \) has covariance matrix \( C \), the linearly transformed random variable \( Ax + b \) has covariance matrix \( ACA^\top \). This is due to the linearity of expectation,
\[ \mathbb{V}(Ax + b) = \mathbb{E} \left[ (Ax + b - \mathbb{E}(Ax + b))(Ax + b - \mathbb{E}(Ax + b))^\top \right] \] (1.46)
\[ = \mathbb{E} \left[ (Ax - \mathbb{E}(Ax))(Ax - \mathbb{E}(Ax))^\top \right] \] (1.47)
\[ = \mathbb{E} \left[ (A(x - \mathbb{E}(x)))(A(x - \mathbb{E}(x)))^\top \right] \] (1.48)
\[ = \mathbb{E} \left[ A(x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top A^\top \right] \] (1.49)
\[ = A \mathbb{E} \left[ (x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top \right] A^\top \] (1.50)
\[ = ACA^\top. \] (1.51)

The correlation matrix \( K \) is the \( d \times d \) matrix with all correlation coefficients. The correlation coefficient is essentially the covariance between two random variables that are standardised to unit variance (and zero mean), which can be achieved by the linear transformation \( D^{-1/2}(x - \mu) \) where \( D \) contains the diagonal elements of \( C \), i.e. \( D = \text{diag}(\mathbb{V}(x_1), \ldots, \mathbb{V}(x_d)) \). We thus have
\[ K = D^{-1/2}CD^{-1/2}. \] (1.52)

By construction, the diagonal elements of \( K \) are all one. Both the correlation matrix and the covariance matrix can be computed from data \( x_1, \ldots, x_n \) by replacing the expectations with sample averages. In Section 1.3.2, we present a compact matrix expression.

If the dimension \( d \) is not too large, we can just display the numerical values of the covariance or correlation matrix. But for large \( d \), this is not helpful. For large dimensions, options are
- visualising \( C \) or \( K \) as an \( d \times d \) image with suitably chosen colour codes,
- summarising the distribution of the variances and correlation coefficients by any of the techniques from the previous section, e.g. histograms or kernel density estimates,
- doing principal component analysis for a more in-depth analysis (see Chapter 2).
Nonlinear relationships

A simple way to measure possible nonlinear relationships between two random variables \( x \) and \( y \) is to compute their covariance or correlation after transforming them nonlinearly, i.e. to compute

\[
\rho(g(x), g(y)) = \frac{\text{cov}(g(x), g(y))}{\sqrt{\text{V}(g(x)) \text{V}(g(y))}}
\]

for some nonlinear function \( g \).

Figure 1.9 shows two examples. In Figure 1.9(a) there is a clear functional relation between \( x \) and \( y \) but the (linear) correlation coefficient is \(-0.15\), wrongly indicating a negative correlation between \( x \) and \( y \). Computing the correlation between the absolute values \(|x|\) and \(|y|\), however, yields a correlation coefficient of 0.93. In Figure 1.9(b), the variance of \( y \) depends on the magnitude of \( x \). The linear correlation is practically zero while the absolute values have a correlation coefficient of 0.68.

Different nonlinearities \( g \) can be used to measure different properties of the data. The absolute value, for example, can be used to measure variance dependencies. And to measure whether \( x \) and \( y \) are monotonically related, one can correlate the ranks of the data points instead of their actual value. The latter quantity is known as Spearman’s rank correlation coefficient.

1.2.2 Graphs

The graphs in Figure 1.9 are known as scatter plots. Each bubble corresponds to a data point. The scatter plot is one of the most used techniques to visualise the distribution of two variables. Colouring the bubbles and changing their size enables the visualisation of further dimensions or class labels.

Other graphs that are being used are two-dimensional extensions of the histogram and the kernel density estimator. For random vectors, we can show these
1.3 Simple preprocessing

Preprocessing refers to various operations that need to be performed in order to prepare the data for further analysis. We here discuss simple methods for outlier detection and data standardisation.

1.3.1 Simple outlier detection

An outlier is a data point that seems unusual compared to others. This is a vague definition, which reflects the various possible causes for outliers.

An outlier can be due to an error in the data gathering stage, for example because a measurement device did not work properly (“bad data”). A data point may, however, also appear unusual because it does not conform to the current assumptions that are made about the data. In the former case, we may omit the corrupted data points from the analysis, while in the latter case, the data points contain valuable information that should not be discarded.

Some bad data points can be spotted by the methods above for describing univariate or bivariate data. If there is a strong difference between the sample mean and sample median, for example, the cause may be a bad data point. Quantiles, histograms, and scatter plots further enable one to spot bad data points. Tukey’s test is a classical method that is based on quartiles and considers data points outside the range

$$[Q_1 - k \text{ IQR}, Q_3 + k \text{ IQR}]$$

(1.54)
as outliers. Typically $k = 1.5$, so that points that fall beyond the whiskers of the boxplot in Figure 1.7 are considered outliers.

1.3.2 Data standardisation

Data standardisation refers classically to normalising the data to have zero (sample) mean and unit (sample) variance. It may, however, also refer to other kinds of transformations to make all variables comparable, for example, transforming the variables to be in $[0, 1]$. Common further transformations that are being used are removing the average value of each single data vector, re-scaling the vector to unit norm, or computing the logarithm of its values. The transformations are often problem dependent.

Denote the “raw” data vector by $\tilde{x}_1, \ldots, \tilde{x}_n$ and the corresponding data matrix by $\tilde{X}$,

$$\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n).$$

(1.55)

We next write the removal of the mean (centring) and scaling to unit variance as matrix operations.
Centring

Let us denote the average of the \( \tilde{x}_i \) by \( m \) and the centred data points by \( x_i \), so that

\[
x_i = \tilde{x}_i - m.
\]  

(1.56)

Let us further collect all centred data points into the matrix \( X \),

\[
X = (x_1, \ldots, x_n).
\]  

(1.57)

We will now see that the centring operation can be written compactly as a matrix multiplication,

\[
X = \tilde{X} H_n, \quad H_n = I_n - \frac{1}{n} 1_n 1_n^\top,
\]  

(1.58)

where \( I_n \) is the \( n \times n \) identity matrix and \( 1_n^\top = (1, 1, \ldots, 1) \) is a vector of ones. The matrix \( H_n \) is called the centring matrix.

This can be seen as follows: From (1.9), we know that we can express the sample mean \( m = (m_1, \ldots, m_d)^\top \) as

\[
m = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i = \tilde{X} \frac{1}{n} 1_n,
\]  

(1.59)

Note that \( m_i \) is the average over all elements in row \( i \) of the matrix \( \tilde{X} \), i.e.

\[
m_i = \frac{1}{n} \sum_{j=1}^{n} (\tilde{X})_{ij},
\]  

(1.60)

where \( (\tilde{X})_{ij} \) denotes the \((ij)\)-th element of \( \tilde{X} \). Since \( X \) is obtained by subtracting vector \( m \) from each column of \( \tilde{X} \), we have

\[
X = (\tilde{x}_1 - m, \ldots, \tilde{x}_n - m) = (\tilde{x}_1, \ldots, \tilde{x}_n) - (m, \ldots, m) = \tilde{X} - (m, \ldots, m)
\]  

(1.61)

The matrix \( (m, \ldots, m) \) has \( n \) copies of \( m \) as its columns. We can write \( (m, \ldots, m) \) as the outer product \( m 1_n^\top \) between \( m \) and \( 1_n \),

\[
m 1_n^\top = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} m_1 & m_1 & \cdots & m_1 \\ m_2 & m_2 & \cdots & m_2 \\ \vdots & \vdots & \ddots & \vdots \\ m_d & m_d & \cdots & m_d \end{pmatrix} = (m, \ldots, m). \]  

(1.62)

With \( m = \tilde{X} \frac{1}{n} 1_n \), we thus obtain

\[
(m, \ldots, m) = \tilde{X} \frac{1}{n} 1_n 1_n^\top
\]  

(1.63)

and hence

\[
X = \tilde{X} - (m, \ldots, m) = \tilde{X} - \tilde{X} \frac{1}{n} 1_n 1_n^\top = \tilde{X} (I_n - \frac{1}{n} 1_n 1_n^\top) = \tilde{X} H_n,
\]  

(1.64)
as claimed in (1.58).

Multiplying $\bar{X}$ with $H_n$ from the right subtracts the average value of each row of $\bar{X}$ from each element in said row, i.e.

$$ (\bar{X}H_n)_{ij} = (\bar{X})_{ij} - \frac{1}{n} \sum_{j=1}^{n} (\bar{X})_{ij}. \quad (1.65) $$

As a side note, multiplying $H_n$ from the left with a column vector $a = (a_1, \ldots, a_n^\top)$ would subtract the average of all $a_i$ from each element of $a$,

$$ H_n a = \begin{pmatrix} a_1 - \bar{a} \\ a_2 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix} \quad \text{and } \bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i, \quad (1.66) $$

and hence, multiplying a matrix with $H_n$ from the left would subtract the average of each column from each column of the matrix. In brief, $H_n$ is a projection matrix that projects vectors on the space orthogonal to $1_n$ (see Section A.6). It satisfies $H_n H_n = H_n$.

**Scaling to unit variance**

The sample covariance matrix $\hat{C}$ for the (raw) data is

$$ \hat{C} = \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - \bar{m})(\bar{x}_i - \bar{m})^\top \quad (1.67) $$

$$ = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \quad (1.68) $$

$$ = \frac{1}{n} \bar{X} \bar{X}^\top \quad (1.69) $$

$$ = \frac{1}{n} \bar{X} H_n \bar{X}^\top. \quad (1.70) $$

Its diagonal elements contains the sample variances. Let us collect them into the diagonal matrix $D$. We can then scale the centred data $x_i$ to unit variance by pre-multiplying them with $D^{-1/2}$. Note that each dimension of the data points $D^{-1/2} x_i$ has now unit variance, but they may still be correlated. In fact, the (sample) covariance matrix of the rescaled data $D^{-1/2} x_i$ is an estimate of the correlation matrix $K$ in (1.52). Finally, we will later often use $C$ to denote not only the covariance matrix but also the sample covariance matrix. The context will make clear which matrix is meant.

**References**


Appendix A

Linear algebra

The material in this chapter is mostly a refresher of some basic results from linear algebra. But it also contains some proofs of results that may be harder to find. The proofs are not examinable.

A.1 Matrices

A $m \times n$ matrix $A$ is a $m \times n$ array of numbers arranged into $m$ rows and $n$ columns. The element at row $i$ and column $j$ is denoted by $a_{ij}$ so that

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}. \quad (A.1)$$

We will sometimes use the indexing notation $(A)_{ij}$ to refer to element $a_{ij}$. The transpose $A^\top$ of a matrix $A$ is the matrix where the entries of $A$ are mirrored at the diagonal, i.e. $(A^\top)_{ij} = (A)_{ji}$. If $A^\top = A$, the matrix is said to be symmetric.

Multiplying a matrix with a scalar produces a matrix where each element is scaled by said scalar, for example

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \ldots & \alpha a_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha a_{m1} & \alpha a_{m2} & \ldots & \alpha a_{mn} \end{bmatrix}. \quad (A.2)$$

Two matrices of the same size can be added together by adding their corresponding elements, for example

$$A + B = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \ldots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \ldots & a_{1n} + b_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \ldots & a_{mn} + b_{mn} \end{bmatrix}. \quad (A.4)$$
If matrix $A$ has size $m \times n$ and matrix $B$ size $n \times p$, the two matrices can be multiplied together. The results is a $m \times p$ matrix $C = AB$ whose elements $(C)_{ij} = c_{ij}$ are given by

$$(C)_{ij} = \sum_{k=1}^{n} (A)_{ik} (B)_{kj}, \quad \text{or, equivalently,} \quad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \quad (A.5)$$

The equations mean that to compute the $(ij)$-th element of $C$, we multiply the elements of the $i$-th row of $A$ with the elements of the $j$-th column of $B$ and sum them all up.

The trace of a $m \times m$ matrix $A$ is the sum of its diagonal elements,

$$\text{trace}(A) = \sum_{i=1}^{m} a_{ii}. \quad (A.6)$$

The trace of $AB$ equals the trace of $BA$: Let $A$ be $m \times n$ and $B$ $n \times m$. We then have

$$\text{trace}(AB) = \sum_{i=1}^{m} (AB)_{ii} = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} a_{ij} b_{ji} \right) = \sum_{j=1}^{n} \sum_{i=1}^{m} b_{ji} a_{ij}, \quad (A.7)$$

which equals $\sum_{j=1}^{n} (BA)_{jj}$ and hence

$$\text{trace}(AB) = \text{trace}(BA) \quad (A.8)$$

as claimed.

### A.2 Vectors

A $n$-dimensional vectors $v$ can be seen as $n \times 1$ matrix. We denote its $i$-th element by $v_i$ or sometimes also by $(v)_i$. By default, $v$ is a column vector, i.e.

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}. \quad (A.9)$$

It’s transpose $v^\top$ is the row vector $(v_1, \ldots, v_n)$. Like matrices, vectors can be scaled, added or multiplied together. The product between a $1 \times n$ vector $u$ and a $n \times 1$ vector $v$ is with (A.5) a number equal to

$$uv = \sum_{i=1}^{n} u_i v_i. \quad (A.10)$$

The inner product or scalar product $u^\top v$ between two $n$ dimensional vectors $u$ and $v$ is

$$u^\top v = \sum_{i=1}^{n} u_i v_i, \quad (A.11)$$
that is, the vector $u$ is first transposed to be row vector after which (A.5) is applied. Importantly, it does not matter whether $u$ or $v$ is transposed, i.e.

$$u^\top v = v^\top u.$$  \hfill (A.12)

The outer product $uv^\top$ between a $m$ dimensional vector $u$ and a $n$ dimensional vector $v$ is a $m \times n$ matrix

$$uv^\top = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} = \begin{pmatrix} u_1 v_1 & u_1 v_2 & \cdots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \cdots & u_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_m v_1 & u_m v_2 & \cdots & u_m v_n \end{pmatrix}. \hfill (A.13)$$

It can be seen that the $(i, j)$-th element of the matrix is equal to $u_i v_j$ in line with (A.5).

Equation (A.5) also tells us that the product between a $m \times n$ matrix $A$ and $n$-dimensional vector $v$ equals a $m$-dimensional vector $v$ with elements $v_i$,

$$v_i = \sum_{j=1}^{n} a_{ij} u_j \quad i = 1, \ldots, m. \hfill (A.14)$$

### A.3 Matrix operations as operations on column vectors

It is often helpful to consider a $m \times n$ matrix $A$ as a collection of $n$ column vectors $a_j$ of dimension $m$ that are arranged next to each other,

$$A = (a_1, \ldots, a_n). \hfill (A.15)$$

Note that the $i$-th element of the $j$-th column of $A$ is $(A)_{ij} = (a_j)_i$.

#### A.3.1 Matrix-vector products

By computing the $i$-th element, we see that $v = Au$ can be written as weighted combination of the column vectors $a_j$,

$$Au = \sum_{j=1}^{n} a_j u_j = \underbrace{\begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix}}_{a_j} u_1 + \cdots + \underbrace{\begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}}_{a_j} u_j + \cdots + \underbrace{\begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix}}_{a_j} u_n, \hfill (A.16)$$

The equation shows that for vectors $u$ that are zero everywhere but in slot $k$, $Au = a_k u_k$, which means that we can “pick” column $k$ of $A$ by multiplication with the $k$ unit vector.
A.3.2 Matrix-matrix products

Products between matrices can also be written in terms of operations on the column vectors. Let $B$ be a $n \times p$ matrix with column vectors $b_i \in \mathbb{R}^n$,

$$B = (b_1, \ldots, b_p).$$ (A.17)

By computing the $(i, j)$-th element, we see that $AB$ can be written as a collection of column vectors $Ab_j$,

$$AB = (Ab_1, \ldots, Ab_p).$$ (A.18)

Indeed, the $i$-th element of the $j$-th column is $(Ab_j)_i$ and

$$(Ab_j)_i = \sum_{k=1}^n (A)_{ik}(b_j)_k = \sum_{k=1}^n (A)_{ik}(B)_{kj},$$ (A.19)

which equals $(AB)_{ij}$.

Assume for a moment that matrix $B$ is zero everywhere but in a $r \times r$ block in the upper left,

$$B = \begin{pmatrix} b_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$ (A.20)

That is, the first $r$ column vectors $b_j$ are zero everywhere but in slot $j$ where they equal $b_j$, i.e. $b_1 = (b_1, 0, \ldots)^\top$, $b_2 = (0, b_2, 0, \ldots)^\top$ and so on, and the remaining column vectors $b_{r+1}, \ldots, b_p$ are all zero. From (A.18) and (A.16), it follows that

$$AB = (b_1a_1, b_2a_2, \ldots, b_ra_r, 0, \ldots, 0).$$ (A.21)

This shows that we can weigh each column vector of the matrix $A$, or set it to zero, by multiplying it with a matrix that is zero everywhere but in the first $r$ diagonal elements.

A.3.3 Outer product representation of a matrix-matrix product

Assume we want to compute the matrix product $AB^\top$ where $A$ is $m \times n$ as before but $B$ is $p \times n$. Let us denote the $n$ columns of $B$ by $b_j \in \mathbb{R}^p$,

$$B = (b_1, \ldots, b_n).$$ (A.22)

From (A.5), we know that

$$(AB^\top)_{ij} = \sum_{k=1}^n (A)_{ik}(B^\top)_{kj} = \sum_{k=1}^n (A)_{ik}(B)_{kj}$$ (A.23)

We now show that $AB^\top$ can also be written as sum of outer products between the column vectors of $A$ and $B$,

$$AB^\top = \sum_{k=1}^n a_kb_k^\top.$$ (A.24)
This identity can be verified by computing the \((i, j)\)-th element of the matrix on the right-hand-side:

\[
\left( \sum_{k=1}^{n} a_k b_k^\top \right)_{i,j} = \sum_{k=1}^{n} (a_k b_k^\top)_{i,j} = \sum_{k=1}^{n} (a_k)_i (b_k)_j. \tag{A.25}
\]

\[
\left( \sum_{k=1}^{n} a_k b_k^\top \right)_{i,j} = \sum_{k=1}^{n} (a_k)_i (b_k)_j. \tag{A.26}
\]

Since \((a_k)_i\) is the \(i\)-th element of the \(k\)-th column of \(A\), we have \((a_k)_i = (A)_{ik}\). For the same reason, \((b_k)_j = (B)_{jk}\), so that \((a_k)_i (b_k)_j = (A)_{ik} (B)_{jk}\) and

\[
\left( \sum_{k=1}^{n} a_k b_k^\top \right)_{i,j} = \sum_{k=1}^{n} (A)_{ik} (B)_{jk}, \tag{A.27}
\]

which equals (A.23) and thus proves the identity in (A.24).

A.4 Orthogonal basis

Two vectors \(u_1 \in \mathbb{R}^n\) and \(u_2 \in \mathbb{R}^n\) are said to be orthogonal if their inner product (scalar product) \(u_1^\top u_2\) is zero. If additionally the vectors are of unit norm, \(||u_i|| = \sqrt{u_i^\top u_i}, \quad i = 1, 2, \tag{A.28}\)

the vectors are said to be orthonormal. A set of \(n\) orthonormal vectors \(u_i \in \mathbb{R}^n\) forms an orthogonal basis of \(\mathbb{R}^n\). This means that any vector \(x \in \mathbb{R}^n\) can be written as a weighted combinations of the \(u_1, \ldots, u_n\),

\[
x = \sum_{i=1}^{n} c_i u_i. \tag{A.29}
\]

The weights \(c_i\) are the coordinates of \(x\) with respect to the basis. Due to the orthogonality of the \(u_i\), the coordinates \(c_i\) can computed via an inner product between the \(u_i\) and \(x\),

\[
c_i = u_i^\top x, \quad i = 1, \ldots, n, \tag{A.30}
\]

We can form a matrix \(U\) by putting all the orthonormal basis vectors next to each other as the columns of the matrix,

\[
U = (u_1, \ldots, u_n). \tag{A.31}
\]

The matrix \(U\) is said to be an orthogonal matrix. Since the vectors \(u_i\) have unit norm and are orthogonal to each other, we have that \(U^\top U = I_n\) where \(I_n\) is the \(n\)-dimensional identity matrix.

Collecting all coordinates \(c_i\) into the vector \(c = (c_1, \ldots, c_n)^\top\), we have with (A.30)

\[
c = U^\top x. \tag{A.32}
\]

With (A.16), we can similarly write (A.29) more compactly as

\[
x = U c. \tag{A.33}
\]

It follows that \(x = U U^\top x\), from where we see that not only \(U^\top U = I_n\) but also \(U U^\top = I_n\) for orthogonal matrices \(U\).
A.5 Subspaces

An orthogonal basis \( u_1, \ldots, u_n \) enables us to represent any vector \( x \in \mathbb{R}^n \) as a weighted combination of the vectors. If we do not have \( n \) orthonormal vectors but only \( k \) of them, e.g. \( u_1, \ldots, u_k \), we cannot represent all \( n \)-dimensional vectors but only those vectors \( z \in \mathbb{R}^n \) that can be written as

\[
z = \sum_{i=1}^{k} a_i u_i, \quad a_i \in \mathbb{R}.
\] (A.34)

This set of vectors is said to be spanned by the \( u_1, \ldots, u_k \) and denoted by \( \text{span}(u_1, \ldots, u_k) \). In other words,

\[
\text{span}(u_1, \ldots, u_k) = \{ z \in \mathbb{R}^n : z = \sum_{i=1}^{k} a_i u_i \}.
\] (A.35)

If \( z_1 \in \text{span}(u_1, \ldots, u_k) \) and \( z_2 \in \text{span}(u_1, \ldots, u_k) \), i.e. if

\[
z_1 = \sum_{i=1}^{k} a_i u_i, \quad z_2 = \sum_{i=1}^{k} b_i u_i,
\] (A.36)

their weighted sum \( \alpha z_1 + \beta z_2 \) equals

\[
\alpha z_1 + \beta z_2 = \sum_{i=1}^{k} \alpha a_i u_i + \sum_{i=1}^{k} \beta b_i u_i = \sum_{i=1}^{k} (\alpha a_i + \beta b_i) u_i
\] (A.37)

and thus belongs to \( \text{span}(u_1, \ldots, u_k) \) as well. This means that the span is closed under addition and scalar multiplication, which makes it a subspace of \( \mathbb{R}^n \). Since any vector \( z \) of \( \text{span}(u_1, \ldots, u_k) \) can be expressed using \( k \) coordinates only, namely the \( u_i^j z, i = 1, \ldots, k \), \( \text{span}(u_1, \ldots, u_k) \) is a \( k \)-dimensional subspace of \( \mathbb{R}^n \).

We now show that any vector \( x \in \mathbb{R}^n \) can be split into a part \( x_\parallel \) that belongs to \( \text{span}(u_1, \ldots, u_k) \) and a part \( x_\perp \) that belongs to \( \text{span}(u_{k+1}, \ldots, u_n) \), the span of the remaining basis vectors \( u_{k+1}, \ldots, u_n \). Since

\[
x = \sum_{j=1}^{n} u_j c_j = \sum_{j=1}^{k} u_j c_j + \sum_{j=k+1}^{n} u_j c_j
\] (A.38)

we have that

\[
x = x_\parallel + x_\perp, \quad x_\parallel = \sum_{j=1}^{k} u_j c_j, \quad x_\perp = \sum_{j=k+1}^{n} u_j c_j.
\] (A.39)

As \( x_\parallel \) is a weighted sum of the \( u_1, \ldots, u_k \), and \( x_\perp \) a weighted sum of the \( u_{k+1}, \ldots, u_n \), the vectors \( x_\parallel \) and \( x_\perp \) are orthogonal to each other. The subspace \( \text{span}(u_{k+1}, \ldots, u_n) \) is said to be orthogonal to \( \text{span}(u_1, \ldots, u_k) \) and is thus also denoted by \( \text{span}(u_1, \ldots, u_k)^\perp \).
A.6 Orthogonal projections

Let us collect the \( k \) vectors \( u_k \) into the \( n \times k \) matrix \( U_k \),

\[
U_k = (u_1, \ldots, u_k).
\] (A.40)

Since the \( u_k \) are orthonormal, \( U_k^\top U_k = I_k \), but, unlike for orthogonal matrices, \( U_k U_k^\top \) is not the identity matrix. We next show that \( U_k U_k^\top x \) equals the part \( x_\parallel \) of \( x \) that belongs to the \( k \)-dimensional subspace \( \text{span}(u_1, \ldots, u_k) \).

This can be most easily seen by writing \( U_k U_k^\top \) as a sum of elementary matrices \( u_i u_i^\top \),

\[
U_k U_k^\top = \sum_{i=1}^{k} u_i u_i^\top,
\] (A.41)

which we can do according to (A.24). Applying \( U_k U_k^\top \) on a vector \( x \) thus gives

\[
U_k U_k^\top x = \sum_{i=1}^{k} u_i u_i^\top x
\] (A.42)

\[
= \sum_{i=1}^{k} u_i^\top x u_i
\] (A.43)

\[
= \sum_{i=1}^{k} u_i^\top \sum_{j=1}^{n} u_j c_j
\] (A.44)

\[
= \sum_{i=1}^{k} u_i c_i
\] (A.45)

\[
= x_\parallel,
\] (A.46)

where we have used that \( u_i^\top u_j \) equals zero unless \( j = i \). The mapping of \( x \) to \( U_k U_k^\top x = x_\parallel \) is called the orthogonal projection of \( x \) onto \( \text{span}(u_1, \ldots, u_k) \). It follows that \( (I_d - U_k U_k^\top) x \) equals \( x_\perp \), and that the matrix \( (I_d - U_k U_k^\top) \) is the orthogonal projection of \( x \) onto \( \text{span}(u_1, \ldots, u_k)^\perp \).

A.7 Singular value decomposition

The singular value decomposition (SVD) of a \( m \times n \) matrix \( A \) is the factorisation of the matrix into the product \( U S V^\top \),

\[
A = U S V^\top,
\] (A.47)

The \( m \times n \) matrix \( S \) is zero everywhere but in the first \( r \) diagonal elements \( (S)_{ii} \) that are positive. We denote the \( (S)_{ii} \) by \( s_i \) so that

\[
S = \begin{pmatrix}
  s_1 & & \\
  & \ddots & \\
  & & s_r \\
 0 & & 0
\end{pmatrix}
\] (A.48)
The diagonal elements are called the singular values of $A$ and are typically ordered so that $s_1 \geq s_2 \geq \cdots \geq s_r$. Matrices $U$ and $V$ are both orthogonal. We denote the column vectors of the two matrices correspondingly by $u_i$ and $v_i$,

$$U = (u_1, \ldots, u_m), \quad V = (v_1, \ldots, v_n).$$ (A.49)

The vectors $u_i$ and $v_i$ form an orthogonal basis for $\mathbb{R}^m$ and $\mathbb{R}^n$, and are called the left-singular vectors and right-singular vectors, respectively. The number $r \leq \min(m,n)$ is called the rank of the matrix $A$.

Due to the structure of the matrix $S$ only the $u_i$ and $v_i$ with $i \leq r$ actually contribute to the factorisation. Indeed, with (A.21), the $m \times n$ matrix $US$ equals

$$US = (s_1 u_1, \ldots, s_r u_r, 0, \ldots, 0).$$ (A.50)

and with (A.24), $USV^\top$ is

$$USV^\top = \sum_{i=1}^{r} s_i u_i v_i^\top + \sum_{i=r+1}^{n} 0v_i^\top$$ (A.51)

so that $A = USV^\top$ is

$$A = \sum_{i=1}^{r} s_i u_i v_i^\top = U_r S_r V_r^\top$$ (A.52)

where

$$U_r = (u_1, \ldots, u_r), \quad S_r = \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_r \end{pmatrix}, \quad V_r = (v_1, \ldots, v_r).$$ (A.53)

This is called the compact, “thin”, or “skinny” SVD of $A$.

### A.8 Eigenvalue decomposition

The eigenvalue decomposition is a factorisation for symmetric matrices. The eigenvalue decomposition of the symmetric $m \times m$ matrix $A$ of rank $r$ is

$$A = U \Lambda U^\top,$$ (A.54)

where $\Lambda$ is a $m \times m$ diagonal matrix with $r$ non-zero elements $\lambda_i$ that we can assume to be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. Note that the $\lambda_i$ may be positive or negative. Matrix $U$ is orthogonal with orthonormal column vectors $u_i$. As for the SVD, the vectors $u_i$ for which $(A)_{ii} = 0$ can actually be ignored so that

$$A = \sum_{i=1}^{r} \lambda_i u_i u_i^\top = U_r \Lambda_r U_r^\top,$$ (A.55)

where

$$U_r = (u_1, \ldots, u_r), \quad \Lambda_r = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}$$ (A.56)
The vectors $u_i$ are called the eigenvectors and the $\lambda_i$ the eigenvalues. It follows from (A.55) that

$$Au_k = \lambda_k u_k,$$  \hspace{1cm} (A.57)

i.e. the matrix $A$ only scales the vectors $u_i$ by their corresponding eigenvalue $\lambda_i$.

## A.9 Positive semi-definite and definite matrices

A symmetric $m \times m$ matrix is called positive semi-definite if all $m$ eigenvalues are non-negative and positive definite if they are all positive. A positive definite matrix has full rank, $r = m$, and the eigenvectors $u_1, \ldots, u_m$ form an orthogonal basis of $\mathbb{R}^m$.

If a matrix $M$ has the singular value decomposition $M = U_r S_r V_r^\top$ as in (A.52), the eigenvalue decomposition of $MM^\top$ is

$$MM^\top = U_r S_r V_r^\top V_r S_r U_r^\top = U_r S_r^2 U_r^\top,$$  \hspace{1cm} (A.58)

on the other hand, the eigenvalue decomposition of $M^\top M$ is

$$M^\top M = V_r S_r U_r^\top U_r S_r V_r^\top = V_r S_r^2 V_r^\top,$$  \hspace{1cm} (A.59)

where in both cases $S_r^2$ refers to the diagonal matrix with elements $s^2_i$. Both $M^\top M$ and $MM^\top$ have the $s^2_i$ as eigenvalues. We see that the eigenvalues are non-negative so that $M^\top M$ and $MM^\top$ are positive semi-definite matrices.

## A.10 Matrix approximations

### A.10.1 Low rank approximation of general matrices

The singular value decomposition allows us to decompose a $m \times n$ matrix $A$ of rank $r$ as

$$A = \sum_{i=1}^{r} s_i u_i v_i^\top = U_r S_r V_r^\top,$$  \hspace{1cm} (A.60)

see (A.52). The $r$ singular values $s_i > 0$ are decreasing. Intuitively, the “later” rank-one matrices $u_i v_i^\top$ with smaller singular values contribute less to $A$ than the “earlier” rank-one matrices with larger singular values. In fact the best approximation $\hat{A}$ of the matrix $A$ by a matrix $\tilde{A}$ of rank $k < r$ is given by the first $k$ terms of the expansion above,

$$\hat{A} = \sum_{i=1}^{k} s_i u_i v_i^\top.$$  \hspace{1cm} (A.61)

This result is unique if and only if $s_k > s_{k+1}$. The result is obtained when the quality of the approximation is measured by the Frobenius norm

$$||A - \hat{A}||_F = \sum_{ij} ((A)_{ij} - (\hat{A})_{ij})^2$$  \hspace{1cm} (A.62)
but also for other matrix norms (e.g. the spectral norm). For the Frobenius norm, the error when approximating $A$ with $\hat{A}$ is the sum of the squares of the remaining singular values

$$\sum_{ij} ((A)_{ij} - (\hat{A})_{ij})^2 = \sum_{i=k+1}^{r} s_i^2.$$  \hfill (A.63)

This result is known as the Eckart–Young–Mirsky theorem and a proof can be found in e.g. (Gentle, 2007, Section 3.10) or (Björck, 2015, Theorem 2.2.11).

A.10.2 Low rank approximation of positive semi-definite matrices

For positive semi-definite matrices, the above approximation based on the singular value decomposition carries over: The best approximation $\hat{A}$ of a positive semi-definite matrix $A$ of rank $r$ by a matrix $\tilde{A}$ of rank $k < r$ is

$$\hat{A} = \sum_{i=1}^{k} \lambda_i u_i u_i^\top.$$  \hfill (A.64)

The smallest approximation error for the Frobenius norm is

$$\|A - \hat{A}\|_F = \sum_{ij=1}^{m} ((A)_{ij} - (\hat{A})_{ij})^2 = \sum_{i=k+1}^{r} \lambda_i^2,$$  \hfill (A.65)

so that $\|A - \tilde{A}\|_F \geq \sum_{i=k+1}^{r} \lambda_i^2$ for other candidates $\tilde{A}$.

A.10.3 Approximating symmetric matrices by positive semi-definite matrices

A rank $r$ symmetric matrix $A$ that is not positive definite has the eigenvalue decomposition

$$A = \sum_{i=1}^{r} \lambda_i u_i u_i^\top,$$  \hfill (A.66)

where some $\lambda_i$ are negative. Let us assume that there are $p \geq 1$ positive eigenvalues and that $\lambda_1 \geq \ldots \geq \lambda_p > 0 > \lambda_{p+1} \geq \ldots \geq \lambda_r$. We would like to determine the positive semi-definite matrix closest to $A$. Measuring closeness by the Frobenius norm, a result by Higham (1988) shows that the closest matrix $\hat{A}$ is obtained by retaining the terms with positive eigenvalues only,

$$\hat{A} = \sum_{i=1}^{p} \lambda_i u_i u_i^\top = \sum_{i=1}^{r} \max(\lambda_i, 0) u_i u_i^\top.$$  \hfill (A.67)

The approximation error is

$$\|A - \hat{A}\|_F = \sum_{i=p+1}^{r} \lambda_i^2,$$  \hfill (A.68)
and matrix $\hat{A}$ has rank $p$.

Following (Higham, 1988), the proof exploits that the Frobenius norm is invariant under rotations, i.e. $||A||_F = ||AU||_F = ||UA||_F$ for any orthogonal matrix $U$. Let $\tilde{A}$ be a positive semi-definite matrix. We then have

\[
||A - \tilde{A}||_F = ||U_r A_r U_r^\top - \tilde{A}||_F \\
= ||U_r^\top U_r A_r U_r^\top U_r - U_r^\top \tilde{A} U_r||_F \\
= ||\Lambda_r - U_r^\top \tilde{A} U_r||_F \\
= \sum_{i=1}^{r} (\lambda_i - b_{ii})^2 + \sum_{i \neq j} b_{ij}^2
\]  

(A.69)

where $b_{ij}$ are the elements of the matrix $B = U_r^\top \tilde{A} U_r$. Because the $b_{ij}^2 \geq 0$, we have

\[
||A - \tilde{A}||_F \geq \sum_{i=1}^{r} (\lambda_i - b_{ii})^2 \\
= \sum_{i=1}^{p} (\lambda_i - b_{ii})^2 + \sum_{i=p+1}^{r} (\lambda_i - b_{ii})^2 \\
\geq \sum_{i=p+1}^{r} (\lambda_i - b_{ii})^2
\]  

(A.70)

(A.71)

(A.72)

Since $b_{ii} \geq 0$ as $\tilde{A}$ is restricted to be positive semi-definite and $\lambda_i < 0$ for $i > p$, we have in the equation above that $\lambda_i - b_{ii} \leq \lambda_i < 0$ and thus $(\lambda_i - b_{ii})^2 \geq \lambda_i^2$. We thus obtain the following lower bound for $||A - \tilde{A}||_F$:

\[
||A - \tilde{A}||_F \geq \sum_{i=p+1}^{r} \lambda_i^2
\]  

(A.73)

(A.74)

(A.75)

A diagonal matrix $B$ with elements $b_{i} = \max(\lambda_i, 0)$ achieves the lower bound. The result in (A.67) now follows from $\tilde{A} = U_r U_r^\top$.

A.10.4 Low rank approximation of symmetric matrices by positive semi-definite matrices

As before let the symmetric matrix $A$ of rank $r$ have $p$ positive eigenvalues,

\[
A = \sum_{i=1}^{r} \lambda_i u_i u_i^\top,
\]  

(A.76)

where $\lambda_1 \geq \ldots \geq \lambda_p > 0 > \lambda_{p+1} \geq \ldots \geq \lambda_r$. Combining (A.67) with (A.64) we show here that the best positive semi-definite approximation of rank $k < p$ is

\[
\hat{A} = \sum_{i=1}^{k} \lambda_i u_i u_i^\top,
\]  

(A.77)

(A.78)
and that the smallest approximation error is
\[ \|A - \tilde{A}\|_F = \sum_{i=k+1}^{r} \lambda_i^2. \] (A.79)

Let \( \tilde{A} \) be a positive semi-definite matrix of rank \( k < p \). As for the proof of (A.64), we write
\[ \|A - \tilde{A}\|_F = \|U_r \Lambda_r U_r^T - \tilde{A}\|_F \] (A.80)
\[ = \|U_r^T U_r \Lambda_r U_r^T U_r - U_r^T \tilde{A} U_r\|_F \] (A.81)
\[ = \|A_r - U_r^T \tilde{A} U_r\|_F \] (A.82)
\[ = \sum_{i=1}^{r} (\lambda_i - b_{ii})^2 + \sum_{i,j=1 \atop i \neq j}^{r} b_{ij}^2 \] (A.83)

where \( b_{ij} = u_i^T \tilde{A} u_j \) are the elements of the matrix \( B = U_r^T \tilde{A} U_r \). Because the \( b_{ij}^2 \geq 0 \), we have
\[ \sum_{i,j=1 \atop i \neq j}^{r} b_{ij}^2 \geq \sum_{i,j=1}^{p} b_{ij}^2 \] (A.84)
and hence
\[ \|A - \tilde{A}\|_F \geq \sum_{i=1}^{r} (\lambda_i - b_{ii})^2 + \sum_{i,j=1 \atop i \neq j}^{p} b_{ij}^2 \] (A.85)
\[ = \sum_{i=1}^{p} (\lambda_i - b_{ii})^2 + \sum_{i,j=1 \atop i \neq j}^{p} b_{ij}^2 + \sum_{i=p+1}^{r} (\lambda_i - b_{ii})^2 \] (A.86)
\[ = \|A_p - U_p^T \tilde{A} U_p\|_F + \sum_{i=p+1}^{r} (\lambda_i - b_{ii})^2 \] (A.87)

As \( \tilde{A} \) is restricted to be positive semi-definite \( b_{ii} \geq 0 \), and since \( \lambda_i < 0 \) for \( i > p \), we have in the equation above that \( \lambda_i - b_{ii} \leq \lambda_i < 0 \) and thus \( (\lambda_i - b_{ii})^2 \geq \lambda_i^2 \).

Hence:
\[ \|A - \tilde{A}\|_F \geq \|A_p - U_p^T \tilde{A} U_p\|_F + \sum_{i=p+1}^{r} \lambda_i^2 \] (A.88)

The matrix \( A_p \) is a positive definite \( p \times p \) matrix, while the matrix \( U_p^T \tilde{A} U_p \) is a \( p \times p \) matrix of rank \( k \). The smallest approximation error of a positive definite matrix by a matrix of lower rank is with (A.65) equal to \( \sum_{i=k+1}^{p} \lambda_i^2 \). We can thus bound \( \|A - \tilde{A}\|_F \) from below by \( \sum_{i=k+1}^{r} \lambda_i^2 \):
\[ \|A - \tilde{A}\|_F \geq \sum_{i=k+1}^{r} \lambda_i^2. \] (A.89)

The matrix \( \hat{A} \) in (A.78) achieves the lower bound which completes the proof.
References

