Reliable estimation of causal conditional entropy from multivariate time-series data

Martin Kamp Dalgaard

Mathematical Engineering, June 2020

Master's Thesis

This thesis is written in LaTeX. Python 3.7.4 has been applied for data processing and to draw graphs.

Mathematical Engineering
Aalborg University
www.en.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Reliable estimation of causal conditional entropy from multivariate time-series data

Theme:
Information theory and k-nearest neighbors

Project Period:
September 2019-June 2020

Participant:
Martin Kamp Dalgaard

Supervisors:
Jan Østergaard
Jesper Møller

Copies: 1

Page numbers: 62

Date of Completion:
June 3, 2020

Abstract:

This thesis considers how to reliably estimate the causal conditional directed information, which describes the flow of information between different sources, and it is computed with estimators based on the k-nearest neighbors due to their enhanced performance in high dimensions compared to other types of estimators. Both well-known estimators and a new and better estimator are derived, which are tested on both synthetic data and actual EEG data. The hypothesis is that the occipital and frontal areas of the brain are known to have a stronger connectivity when the eyes are closed compared to when they are opened. Therefore, the causal conditional directed information is computed between these areas of the brain when the eyes are both opened and closed, and the differences between these computations are then assessed. The results show that there are only minor differences when the eyes are closed compared to when they are opened, and further studies are required.

Contents

Preface

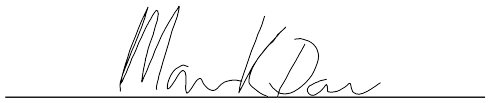
This is a Master's thesis at the Master program in Mathematical Engineering at Aalborg University, and it deals with how to reliably estimate the causal conditional directed information from estimates of the entropy with estimators based on the k-nearest neighbors.

Reading this thesis requires knowledge of basic probability theory, random processes, and stationarity, but additional definitions and results can however be found in Appendix A.

During the project, a library with Python code has been developed, which is used when performing the tests and creating the figures that are described and shown in this thesis. In the spirit of openness, the library as well as the scripts that performs the tests and creates the figures are publicly available at GitHub at this link. Furthermore, `pickle`-files with the outputs from the tests that are used to create the figures in the thesis are also available on GitHub.

The author wishes to thank Jan Østergaard (Department of Electronic Systems), Jesper Møller (Department of Mathematical Sciences), and Payam Shahsavari Baboukani (Department of Electronic Systems) for their guidance during the project.

Aalborg University, June 3, 2020



Martin Kamp Dalgaard
mkda15@student.aau.dk

1 | Introduction

The human brain is a complex network, and it is still difficult to understand how the information is being processed – as opposed to e.g. a digital computer, where each component is known to only perform specific operations on the information.

The mathematician and computer scientist Alan Turing had an idea that information processing can always be decomposed into processes of information storage, transfer, and modification, which are also the processes performed through the hard disk, the CPU, and system buses on a digital computer, respectively [21, pp. V-VI]. Quantifying the information transfer through directed information measures is of particular interest with respect to both neuroscience, complex systems theory in general [21, p. VI], and this thesis.

Measuring the information transfer can describe directed interactions and interdependencies in the brain and whether there is a causal relationship between these interactions [21, p. VI]. Different measures of information transfer are the transfer entropy and directed information [21, pp. 3,28], and a related measure is the *causal conditional directed information* (CCDI) [1, p. 11], which – as the name suggests – considers the causal relationship between the input variables.

As described in Chapter 2, the CCDI can be expressed in terms of the conditional mutual information, which in turn can be expressed through joint entropies. The problem of estimating this measure therefore simplifies to estimating the joint entropies, which can be done through an estimator based on the k-nearest neighbors (kNNs). These kNN-based estimators are popular, computionally efficient, can be asymptotic unbiased, and can outperform e.g. kernel-based estimators in higher dimensions [11, p. 1] [18, p. 304] [25, p. 33], which is required when considering the CCDI. However, kNN-based estimators are not accurate if large correlations are present in the data [11, p. 1].

In practice, the implemented estimators may need to be used in a system which computes the CCDI sequentially (e.g. when new data are available), and it is therefore necessary for the estimators to have a low computation time. Furthermore, it may also be possible to compute the CCDI without using all the available data again every time the CCDI needs to be computed (which would be time-consuming).

The so-called occipital dominant rhytm (referred to as 'alpha waves') can be used to detect "the subject's level of stress, concentration, relaxation or mental load" [2], and researh suggests that there is a higher alpha activity when the eyes are closed [7] [19]. Furthermore, research also suggests that there is a big difference in the connectivity between the occipital and frontal areas of the brain when the eyes are closed

compared to when they are opened [20].

In this thesis, electroencephalographic (EEG) measurements, which are measurements of the electrical activity in the brain, are used. The EEG measurements include sessions with both opened and closed eyes, and the goal is to reliably estimate and use the CCDI to assert the differences between the occipital and frontal areas when the eyes are opened and closed.

1.1 Overview and scope of the thesis

The research question of the thesis is introduced in Section 1.2, and delimitations of the thesis are introduced in Section 1.3. Chapter 2 introduces important elements of information theory such as the mathematical definitions of entropy, mutual information, directed information, and causal conditional directed information. In Chapter 3, both known and new kNN-based estimators are derived. In Chapter 4, the estimators are used on both synthetic data and EEG data to compute the entropy (which is compared to the analytic values of the entropies) and the CCDI when the eyes are opened and closed, respectively. The findings in the thesis are discussed and concluded in Chapters 5 and 6, and further studies are considered in Chapter 7.

In general, the scope of this thesis is the reproducibility of the research on kNN-based estimators. In this aspect, the theory behind estimators from different sources has been understood and described, and these estimators have then been implemented in this thesis. Furthermore, these considerations also made it possible to derive a new estimator. The implemented estimators are firstly measured on their absolute errors when compared to the analytical values of the entropies and are secondly used to compute the CCDI when the eyes are both opened and closed, which are compared to each other in order to either support or refute previous research on information flow in the brain. Finally, it is also considered how the CCDI can be computed sequentially such that the computation time is reduced.

1.2 Research question

As previously described, it is unclear how information flows through the brain but it can be quantified through a directed information measure, which in this thesis is the CCDI, and which in turn can be estimated with kNN-based estimators. The CCDI is computed from actual EEG measurements of sessions with both opened and closed eyes in order to deduce possible connections in the brain, which are hypothesized to be stronger when the eyes are closed compared to when they are opened. However, this requires that the used estimator is able to both reliably and quickly estimate the entropy. The research question in this thesis is:

How can kNN-based entropy estimators be used to estimate the causal conditional directed information measure, which quantifies the information flow between different parts of the brain?

The following study questions are considered in order to be able to answer this:

1. What is the underlying theory of the directed information measure?
2. How can the CCDI be computed sequentially without using all of the available data?
3. How can well-known kNN-based entropy estimators be derived?
4. Is it possible to derive an improved estimator?
5. How can the implemented estimators be tested on both synthetic and actual data?

1.3 Delimitations

This thesis is firstly delimited to only consider kNN-based estimators since they as previously described can outperform e.g. kernel-based estimators in higher dimensions. Secondly, EEG data are primarily considered (apart from synthetic data, which the estimators are initially tested on) since the information flow is the main interest in this thesis. Finally, the CCDI has been chosen as the directed information measure in this thesis from various other measures because it considers the *causal* directed information, which is important when considering the information flow in the brain.

2 | Information theory

In this chapter different elements of information theory are considered. General definitions are firstly considered in Section 2.1, and the concepts of causality and directed information are considered in Section 2.2. The chapter is primarily inspired by [4], [21], and [1].

The random variables considered in this chapter are generally continuous (but the definitions for discrete random variables are analoguous). Furthermore, the entropy of a continuous random variable is usually referred to as *the differential entropy* in order to distinguish it from the entropy of a discrete random variable but for convenience it is just referred to as *the entropy* in this thesis.

The notation used in the following chapters is introduced here. In general, variables are written in bold font when they are vectors. Random variables are denoted by e.g. $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, and several random variables, which form a random process, are indexed as e.g. $\mathbf{X}_i$, $i = 1, \ldots, n$, and they may each produce a realization $\mathbf{x}_i$. Furthermore, the dimension of the random variables is always denoted by $d \in \mathbb{N}$.

2.1 General definitions

In this section, general definitions from information theory are considered. The definition of a random process is firstly considered in Definition 2.1.1.

Definition 2.1.1 *A random process is defined as* $\mathbf{X}^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$*, where* $\mathbf{X}_i \in \mathbb{R}^d$ *is a random variable.*

The entropy in general is defined in Definition 2.1.2 [4, p. 243].

Definition 2.1.2 *The entropy of a continuous random variable* $\mathbf{X} \in \mathbb{R}^d$ *with pdf* $f(\mathbf{x})$ *is defined by*

$$h(\mathbf{X}) = -\int_{\mathbb{R}^d} f(\mathbf{x}) \ln\left(f(\mathbf{x})\right) \mathrm{d}\mathbf{x},$$

where it is assumed that both the pdf $f(\mathbf{x})$ *and the integral exist.*

Remark 2.1 The entropy of $\mathbf{X}$ does not depend on $\mathbf{X}$ but on the pdf of $\mathbf{X}$, $f(\mathbf{x})$. However, the entropy of $\mathbf{X}$ is denoted by $h(\mathbf{X})$ in this thesis due to ease of notation later and because it is the convention used in e.g. [4]. †

In general, the entropy is a measure of the uncertainty of a random variable [4, p. 13], which is examplified through Examples 2.1 and 2.2.

Example 2.1 (Entropy of a normally distributed random variable)
Let $\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_n]^\mathsf{T}$, $\mathbf{X}_i \in \mathbb{R}^d$, be a vector of n random variables with a multivariate normal distribution as defined in Definition A.2.4. The entropy of $\mathbf{X}$ is then:

$$\begin{aligned}
h(\mathbf{X}) &= -\int f(\mathbf{x}) \ln(f(\mathbf{x}))\,\mathrm{d}\mathbf{x} \\
&= -\int f(\mathbf{x}) \ln\left(\frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right)\mathrm{d}\mathbf{x} \\
&= -\int f(\mathbf{x}) \left(\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) - \ln\left((2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}\right)\right)\mathrm{d}\mathbf{x} \\
&= \left(\frac{1}{2}\mathbb{E}\left[\sum_{i,j=1}^{n}(\mathbf{x}_i-\mu_i)\Sigma_{ij}^{-1}(\mathbf{x}_j-\mu_j)\right] + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right)\right)\int f(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\sum_{i,j=1}^{n}\mathbb{E}\left[(\mathbf{x}_j-\mu_j)(\mathbf{x}_i-\mu_i)\right]\Sigma_{ij}^{-1} + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right) \\
&= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\Sigma_{ji}\Sigma_{ij}^{-1} + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right) \\
&= \frac{1}{2}\sum_{j=1}^{n}\left(\Sigma\Sigma^{-1}\right)_{jj} + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right) \\
&= \frac{1}{2}\sum_{j=1}^{n}\mathbf{I}_{jj} + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right) \\
&= \frac{n}{2} + \frac{1}{2}\ln\left((2\pi)^n|\mathbf{\Sigma}|\right) \\
&= \frac{1}{2}\ln\left((2\pi\exp(1))^n|\mathbf{\Sigma}|\right),
\end{aligned}$$

where the following have been used: the definition of the expected value of $\mathbf{X}$ (see (A.3)), that $\int f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$ (see (A.1)), that $\mathbb{E}\left[(\mathbf{x}_j-\mu_j)(\mathbf{x}_i-\mu_i)\right] = \Sigma_{ji}$ (see Definition A.2.4), that $\mathbf{\Sigma}$ is invertible, which means that $\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{I}_n$ (the $n \times n$ identity matrix), and some logarithm rules:

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y), \quad \ln\left(x^c\right) = c\ln(x).$$

Therefore, if the determinant $|\mathbf{\Sigma}|$ of the covariance matrix is small, the uncertainty of $\mathbf{X}$ is small, which means that the entropy is also small.
Note also that for a univariate normal distribution, $X \sim \mathcal{N}(\mu, \sigma^2)$ as defined in Definition A.2.3, the entropy is given by $\frac{1}{2}\ln\left(2\pi\exp(1)\sigma^2\right)$, where σ^2 is the variance.

Example 2.2 (Entropy of a uniformly distributed random variable)
Let $\mathbf{X} \in \mathbb{R}^d$ be a uniformly distributed random variable on a hyperrectangle $R_{\mathbf{a},\mathbf{b}}$ whose vertices' lowest and highest coordinates in each dimension are defined by vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ (a_1 and b_1 are e.g. the vertices' coordinates on the first axis). The pdf of $\mathbf{X}$ is

$$f(\mathbf{x}) = \frac{1}{V(R_{\mathbf{a},\mathbf{b}})} \qquad V(R_{\mathbf{a},\mathbf{b}}) = |a_1 - b_1| \cdot |a_2 - b_2| \cdots |a_d - b_d|,$$

where $V(R_{\mathbf{a},\mathbf{b}})$ is the volume of $R_{\mathbf{a},\mathbf{b}}$. The entropy of $\mathbf{X}$ is then

$$\begin{aligned} h(\mathbf{X}) &= -\int_{R_{\mathbf{a},\mathbf{b}}} f(\mathbf{x}) \ln(f(\mathbf{x}))\, \mathrm{d}\mathbf{x} \\ &= -\frac{1}{V(R_{\mathbf{a},\mathbf{b}})} \ln\left(\frac{1}{V(R_{\mathbf{a},\mathbf{b}})}\right) V(R_{\mathbf{a},\mathbf{b}}) = \ln(V(R_{\mathbf{a},\mathbf{b}})). \end{aligned}$$

■

The entropy of a set of random variables is known as the joint entropy and is defined in Definition 2.1.3 [4, p. 249].

Definition 2.1.3 *The joint entropy of a random process $\mathbf{X}^n$ of n random variables with $\mathbf{X}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, and joint density $f(\mathbf{x}^n)$ is defined as*

$$h(\mathbf{X}^n) = -\int f(\mathbf{x}^n) \ln\left(f(\mathbf{x}^n)\right) \mathrm{d}\mathbf{x}^n.$$

The conditional entropy is defined in Definition 2.1.4 [4, p. 249].

Definition 2.1.4 *The conditional entropy $h(\mathbf{X}|\mathbf{Y})$ of $\mathbf{X}$ given $\mathbf{Y}$ is defined as*

$$h(\mathbf{X} \,|\, \mathbf{Y}) = -\int\int f(\mathbf{x}, \mathbf{y}) \ln(f(\mathbf{x}|\mathbf{y}))\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y},$$

where $f(\mathbf{x}, \mathbf{y})$ is the joint density of $\mathbf{X}, \mathbf{Y}$ and $f(\mathbf{x}|\mathbf{y})$ is the conditional density of $\mathbf{X}$ given $\mathbf{Y}$.

Since $f(\mathbf{x}|\mathbf{y}) = \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{y})}$, the conditional entropy can also be written as:

$$\begin{aligned} h(\mathbf{X} \,|\, \mathbf{Y}) &= -\int\int f(\mathbf{x}, \mathbf{y}) \ln\left(\frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{y})}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int\int f(\mathbf{x}, \mathbf{y}) \left(\ln(f(\mathbf{x}, \mathbf{y})) - \ln(f(\mathbf{y}))\right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int\int f(\mathbf{x}, \mathbf{y}) \ln(f(\mathbf{x}, \mathbf{y})) - f(\mathbf{x}, \mathbf{y}) \ln(f(\mathbf{y}))\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

$$
\begin{aligned}
&= -\int\int f(\mathbf{x},\mathbf{y}) \ln\left(f(\mathbf{x},\mathbf{y})\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \int f(\mathbf{y}) \ln\left(f(\mathbf{y})\right) \mathrm{d}\mathbf{y} \\
&= h(\mathbf{X},\mathbf{Y}) - h(\mathbf{Y}), \qquad (2.1)
\end{aligned}
$$

where $\int\int f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} = \int f(\mathbf{y})\,\mathrm{d}\mathbf{y}$ [14, p. 162].

The joint entropy of n random variables as in Definition 2.1.3 can also be written as a sum of conditional entropies, which is called the chain rule for entropies [4, p. 22]

$$
h(\mathbf{X}^n) = \sum_{i=1}^{n} h(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}), \qquad (2.2)
$$

which follows from repeatedly using (2.1) as expressions of both the joint entropy and conditional entropy [4, p. 22]

$$
\begin{aligned}
h(\mathbf{X}_1,\mathbf{X}_2) &= h(\mathbf{X}_1) + h(\mathbf{X}_2 \,|\, \mathbf{X}_1) \\
h(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3) &= h(\mathbf{X}_1) + h(\mathbf{X}_2,\mathbf{X}_3 \,|\, \mathbf{X}_1) \\
&= h(\mathbf{X}_1) + h(\mathbf{X}_2 \,|\, \mathbf{X}_1) + h(\mathbf{X}_3 \,|\, \mathbf{X}_2,\mathbf{X}_1) \\
&\vdots \\
h(\mathbf{X}^n) &= h(\mathbf{X}_1) + h(\mathbf{X}_2 \,|\, \mathbf{X}_1) + \cdots + h(\mathbf{X}_n \,|\, \mathbf{X}_{n-1},\ldots,\mathbf{X}_1) \\
&= \sum_{i=1}^{n} h\left(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}\right).
\end{aligned}
$$

The mutual information is a measure of the amount of information that one random variable contains about another random variable [4, p. 19] and is defined in Definition 2.1.5.

Definition 2.1.5 *The mutual information $I(\mathbf{X},\mathbf{Y})$ between two random variables $\mathbf{X},\mathbf{Y}$ with joint density $f(\mathbf{x},\mathbf{y})$ is defined as*

$$
I(\mathbf{X};\mathbf{Y}) = \int\int f(\mathbf{x},\mathbf{y}) \ln\left(\frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})f(\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}.
$$

Note that due to Definition 2.1.5, the mutual information is symmetric, i.e. $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{Y};\mathbf{X})$.

The mutual information can be rewritten as:

$$
\begin{aligned}
I(\mathbf{X};\mathbf{Y}) &= \int\int f(\mathbf{x},\mathbf{y}) \ln\left(\frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})f(\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= \int\int f(\mathbf{x},\mathbf{y}) \ln\left(\frac{f(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= -\int\int f(\mathbf{x},\mathbf{y}) \ln(f(\mathbf{x}))\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \int\int f(\mathbf{x},\mathbf{y}) \ln(f(\mathbf{x}|\mathbf{y}))\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= h(\mathbf{X}) - h(\mathbf{X} \,|\, \mathbf{Y}). \qquad (2.3)
\end{aligned}
$$

The conditional entropy $h(\mathbf{X}|\mathbf{Y})$ can not be larger than the entropy $h(\mathbf{X})$ since knowing $\mathbf{Y}$ can not increase the uncertainty of $\mathbf{X}$, and hence $I(\mathbf{X};\mathbf{Y}) \geq 0$ according to (2.3).

The relationship between the entropies, joint entropies, and mutual information of two random variables $\mathbf{X}, \mathbf{Y}$ is shown in Figure 2.1. The relationships in (2.1) and (2.3) can e.g. be seen in the figure. The figure also shows that the mutual information is the reduction in uncertainty of $\mathbf{X}$ due to the knowledge of $\mathbf{Y}$; if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathbf{Y}$ says nothing about $\mathbf{X}$, which means that $h(\mathbf{X}|\mathbf{Y}) = h(\mathbf{X})$, and then $I(\mathbf{X};\mathbf{Y}) = h(\mathbf{X}) - h(\mathbf{X}) = 0$.

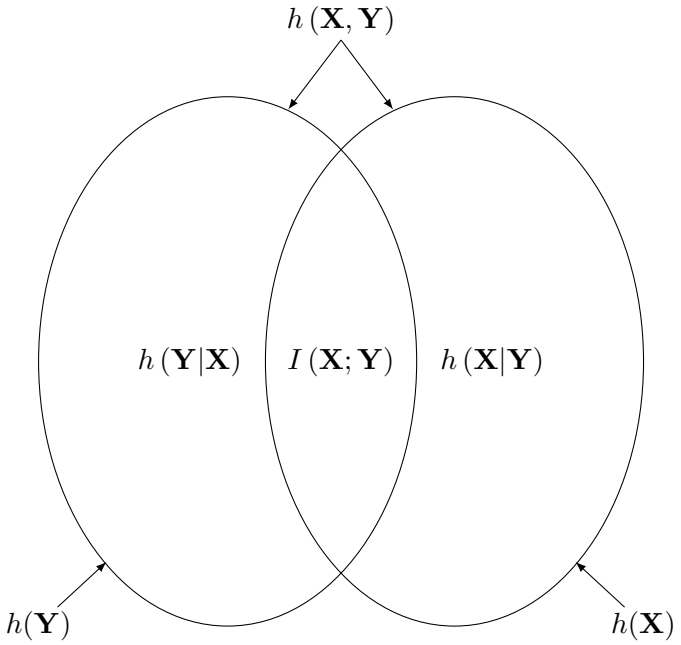


Figure 2.1: This figure is inspired by [4, p. 22].

The conditional mutual information is defined in Definition 2.1.6 [4, p. 23].

Definition 2.1.6 *The conditional mutual information of the random variables* $\mathbf{X}, \mathbf{Y}$ *given* $\mathbf{Z}$ *is defined as*

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = h(\mathbf{X}|\mathbf{Z}) - h(\mathbf{X}|\mathbf{Y},\mathbf{Z}).$$

There is also a chain rule for the mutual information between a random process $\mathbf{X}^n$ of n random variables and a random variable $\mathbf{Y}$. It says that [4, p. 24]

$$I(\mathbf{X}^n;\mathbf{Y}) = \sum_{i=1}^{n} I(\mathbf{X}_i;\mathbf{Y} \mid \mathbf{X}^{i-1}),$$

which follows from applying the result in (2.3), the chain rule for entropies in (2.2), and Definition 2.1.6 such that [4, p. 24]

$$\begin{aligned} I(\mathbf{X}^n;\mathbf{Y}) &= h(\mathbf{X}^n) - h(\mathbf{X}^n \,|\, \mathbf{Y}) \\ &= \sum_{i=1}^{n} h\left(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}\right) - \sum_{i=1}^{n} h\left(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}, \mathbf{Y}\right) \\ &= \sum_{i=1}^{n} I(\mathbf{X}_i;\mathbf{Y} \,|\, \mathbf{X}^{i-1}). \end{aligned}$$

By the same argument, one can obtain a chain rule for the mutual information between two random processes $\mathbf{X}^n$ and $\mathbf{Y}^n$

$$I(\mathbf{X}^n;\mathbf{Y}^n) = \sum_{i=1}^{n} I(\mathbf{X}_i;\mathbf{Y}^n \,|\, \mathbf{X}^{i-1}). \tag{2.4}$$

Section 2.2 considers how these definitions can be applied in causality and directed information.

2.2 Causality and directed information

The first concept considered in this section is Markov chains because it is relevant for the other concepts described in this section.

Different kinds of dependence between random variables in a random process as defined in Definition 2.1.1 are possible. In a Markov chain, each random variable only depends on the previous random variable and is conditionally independent of all the other preceding random variables [4, p. 72]. This is defined in Definition 2.2.1.

Definition 2.2.1 *A random process* $\mathbf{X}^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ *is a Markov chain if*

$$P(\mathbf{X}_i = \mathbf{x}_i \,|\, \mathbf{X}_{i-1} = \mathbf{x}_{i-1}, \ldots, \mathbf{X}_1 = \mathbf{x}_1) = P(\mathbf{X}_i = \mathbf{x}_i \,|\, \mathbf{X}_{i-1} = \mathbf{x}_{i-1}), \quad i = 1, \ldots, n,$$

where $\mathbf{x}_i$ *is a realization of* $\mathbf{X}_i$.

How random variables form a Markov chain is defined in Definition 2.2.2.

Definition 2.2.2 *Three random variables* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *are said to form a Markov chain* $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ *if the joint pdf can be written as*

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) f(\mathbf{y}|\mathbf{x}) f(\mathbf{z}|\mathbf{y}).$$

The two primary concepts described in this section are causality and directed information, which each are described below and combined afterwards.

The concept of Granger causality is inspired by [1, p. 9]. Two random processes

$\mathbf{X}^t$ and $\mathbf{Y}^t$ are considered, where $t \in \mathbb{N}$ denotes a point in time. The random variable $\mathbf{Y}_t$ is said to be 'Granger caused' by $\mathbf{X}_t$ if the prediction of $\mathbf{Y}_t$ is improved when the past of both $\mathbf{Y}_t$ and $\mathbf{X}_t$ are considered. This means that $\mathbf{X}_t$ does not cause $\mathbf{Y}_t$ if and only if $P\left(\mathbf{Y}_t \,|\, \mathbf{Y}^{t-1}, \mathbf{X}^{t-1}\right) = P\left(\mathbf{Y}_t \,|\, \mathbf{Y}^{t-1}\right)$, which can be written as the Markov chain $\mathbf{X}^{t-1} \to \mathbf{Y}^{t-1} \to \mathbf{Y}^t$.

The directed information is defined in Definition 2.2.3 as the sum of conditional mutual informations [21, p. 28].

Definition 2.2.3 *The directed information is defined as*

$$I\left(\mathbf{X}^n \to \mathbf{Y}^n\right) = \sum_{i=1}^{n} I\left(\mathbf{X}^i; \mathbf{Y}_i \,\middle|\, \mathbf{Y}^{i-1}\right),$$

where $\mathbf{Y}^0 \triangleq \emptyset$.

The mutual information and directed information can also be written as follows by using the chain rules for entropies and mutual information in (2.2) and (2.4), respectively, and Definition 2.1.6 of the conditional mutual information

$$\begin{aligned}
I(\mathbf{X}^n; \mathbf{Y}^n) &= I(\mathbf{Y}^n; \mathbf{X}^n) \\
&= \sum_{i=1}^{n} I\left(\mathbf{Y}_i; \mathbf{X}^n \,|\, \mathbf{Y}^{i-1}\right) \\
&= \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{Y}^{i-1}\right) - \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{X}^n, \mathbf{Y}^{i-1}\right) \\
&= h(\mathbf{Y}^n) - \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{X}^n, \mathbf{Y}^{i-1}\right), \\
I(\mathbf{X}^n \to \mathbf{Y}^n) &= \sum_{i=1}^{n} I\left(\mathbf{X}^i; \mathbf{Y}_i \,\middle|\, \mathbf{Y}^{i-1}\right) \\
&= \sum_{i=1}^{n} I\left(\mathbf{Y}_i; \mathbf{X}^i \,\middle|\, \mathbf{Y}^{i-1}\right) \\
&= \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{Y}^{i-1}\right) - \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{X}^i, \mathbf{Y}^{i-1}\right) \\
&= h(\mathbf{Y}^n) - \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{X}^i, \mathbf{Y}^{i-1}\right).
\end{aligned}$$

The only difference between these two expressions is the conditioning, which is the entire random process $\mathbf{X}^n$ for the mutual information and only $\mathbf{X}^i$ for the directed information. The latter has been suggested as *causal conditioning* [1, p. 11].

The concepts of causality and directed information can be combined to the concept of *causal conditional directed information* (CCDI), which is defined as [1, p. 11]

$$\begin{aligned} I\left(\mathbf{X}^n \to \mathbf{Y}^n \,\middle\|\, \mathbf{Z}^n\right) &= \sum_{i=1}^{n} I\left(\mathbf{Y}_i; \mathbf{X}^i \,|\, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right) \\ &= \sum_{i=1}^{n} h\left(\mathbf{Y}_i \,|\, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right) - h\left(\mathbf{Y}_i \,|\, \mathbf{X}^i, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right) \\ &= \sum_{i=1}^{n} h\left(\mathbf{Y}_i, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right) - h\left(\mathbf{Y}^{i-1}, \mathbf{Z}^i\right) \\ &\quad - \left(h\left(\mathbf{X}^i, \mathbf{Y}_i, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right) - h\left(\mathbf{X}^i, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right)\right) \\ &= \sum_{i=1}^{n} h\left(\mathbf{Y}^i, \mathbf{Z}^i\right) - h\left(\mathbf{Y}^{i-1}, \mathbf{Z}^i\right) \\ &\quad - \left(h\left(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{Z}^i\right) - h\left(\mathbf{X}^i, \mathbf{Y}^{i-1}, \mathbf{Z}^i\right)\right), \end{aligned} \tag{2.5}$$

where the second equality follows from Definition 2.1.6 and the third equality follows from the derivation in (2.1).

As the name and notation suggests, the CCDI measures the directed information from $\mathbf{X}^n$ to $\mathbf{Y}^n$, where $\mathbf{Z}^n$ has been causally observed. The latter is denoted by the two vertical lines as opposed to only a single vertical line, which means that it is given (or observed) [1, p. 11].

The expression of the CCDI in (2.5) shows that it can be computed from a sum of joint entropies. This means that the key to reliably estimating the CCDI (which is the goal in order to quantify the connections in the human brain as described in Chapter 1) is to reliably estimate these joint entropies.

However, in each term in the sum in (2.5) the entire pasts of the random processes $\mathbf{X}^i$, $\mathbf{Y}^i$, $\mathbf{Z}^i$ are used, which may be unnecessary if they e.g. form a Markov chain, and it is also computationally inefficient if n is large and the CCDI needs to be computed sequentially when new observations become available. Instead, each of the random variables, e.g. $\mathbf{X}_i$, is assumed to only be dependent on the previous l random variables, which is denoted by $\mathbf{X}^{i-l:i}$. This is expressed in (2.6):

$$\begin{aligned} I\left(\mathbf{X}^n \to \mathbf{Y}^n \,\middle\|\, \mathbf{Z}^n\right) &= \sum_{i=l}^{n} h\left(\mathbf{Y}^{i-l:i}, \mathbf{Z}^{i-l:i}\right) - h\left(\mathbf{Y}^{i-1-l:i-1}, \mathbf{Z}^{i-l:i}\right) \\ &\quad - \left(h\left(\mathbf{X}^{i-l:i}, \mathbf{Y}^{i-l:i}, \mathbf{Z}^{i-l:i}\right) - h\left(\mathbf{X}^{i-l:i}, \mathbf{Y}^{i-1-l:i-1}, \mathbf{Z}^{i-l:i}\right)\right). \end{aligned} \tag{2.6}$$

Note that (2.6) is similar to (2.5) but only includes the last l random variables of each random process in each term.

3 | Entropy estimation through k-nearest neighbors

The entropy $h(\mathbf{X})$ as defined in Definition 2.1.2 can only be computed if the pdf $f(\mathbf{x})$ is known, and even if that is the case, one also needs to evaluate the integral, which may be extremely difficult, especially in high dimensions. Instead, the idea is to estimate the entropy through the k-nearest neighbors (kNNs).

Section 3.1 describes the general setup considered in this chapter and derives an initial basic estimator based on this setup. This estimator turns out to be asymptotically biased, which is shown in Section 3.2, and an asymptotically unbiased estimator is derived following this result. In Section 3.3 the foundation of the derivation of other estimators, which assume a constant density, is made, but it turns out that numerical integration is necessary to estimate a key result, which is performed in Section 3.3.1. The estimators based on the results described in Section 3.3 are described in Section 3.4.

3.1 General setup and initial estimator

This section is inspired by [18] [3]. General concepts, which are used in the remainder of this thesis, are firstly defined.

Definition 3.1.1 *An open ball $\mathcal{B}(\mathbf{x}, r)$ with center $\mathbf{x} \in \mathbb{R}^d$ and radius $r > 0$ is defined as*

$$\mathcal{B}(\mathbf{x}, r) = \left\{ \mathbf{y} \in \mathbb{R}^d : \ \|\mathbf{x} - \mathbf{y}\|_2 < r \right\},$$

where $\|\cdot\|_2$ is the Euclidean norm. Note that the term "open" means that boundary points of the ball are not included.

The volume of the ball $\mathcal{B}(\mathbf{x}, r)$ is denoted by V_r and can be shown to be [12, eq. 5.19.4]

$$V(\mathcal{B}(\mathbf{x}, r)) \triangleq V_r = \frac{\pi^{d/2} r^d}{\Gamma(d/2 + 1)}. \tag{3.1}$$

In general, n iid random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\mathbf{X}_i \in \mathbb{R}^d$, with density $f(\cdot)$ and realizations (also referred to as samples) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are considered. In the remainder of this thesis, the ball $\mathcal{B}(\mathbf{x}_i, r_i)$ is restricted to only contain the k-nearest neighbors of

its center $\mathbf{x}_i$, where $k \leq n-1$ is a positive integer and the radius r_i is the Euclidean distance $r(\cdot,\cdot)$ between $\mathbf{x}_i$ and its kNN, $\mathbf{x}_{i,\text{kNN}}$

$$r_i = r(\mathbf{x}_i, \mathbf{x}^n \setminus \{\mathbf{x}_i\}) = \|\mathbf{x}_i - \mathbf{x}_{i,\text{kNN}}\|_2, \quad \mathbf{x}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}.$$

Throughout the chapter, a specific sample is often given, and it is typically said to be $\mathbf{x}_1$ since the random variables are iid. Therefore, the ball $\mathcal{B}(\mathbf{x}_1, r_1)$ is typically considered. Let this ball be surrounded by an annulus of a small width $\mathrm{d}r_1$ such that the kNN of $\mathbf{x}_1$ lies on the boundary of $\mathcal{B}(\mathbf{x}_1, r_1)$. This is depicted in Figure 3.1.

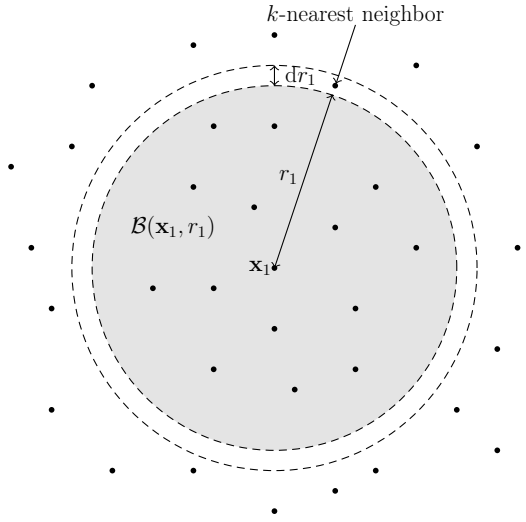


Figure 3.1: The grey area is the ball $\mathcal{B}(\mathbf{x}_1, r_1)$ with center $\mathbf{x}_1$ and radius r_1. The ball is surrounded by an annulus of width $\mathrm{d}r_1$. Note that there is k samples inside $\mathcal{B}(\mathbf{x}_1, r_1)$ (including $\mathbf{x}_1$), one sample in the annulus, and $n-k-1$ samples outside. This figure is inspired by [11, p. 2].

Another general concept is the conditional probability $F(r|\mathbf{x}_1)$ as defined in Definition 3.1.2.

Definition 3.1.2 *The conditional probability that a random variable $\mathbf{X}_2$ lies inside the ball $\mathcal{B}(\mathbf{x}_1, r)$, where $\mathbf{x}_1$ is given, is denoted by $F(r|\mathbf{x}_1)$:*

$$F(r|\mathbf{x}_1) = P(\mathbf{X}_2 \in \mathcal{B}(\mathbf{x}_1, r) \,|\, \mathbf{X}_1 = \mathbf{x}_1) = \int_{\|\mathbf{x}-\mathbf{x}_1\|<r} f(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad r > 0.$$

For small k and large n, the density $f(\mathbf{x}_2)$ of a sample $\mathbf{x}_2$ inside $\mathcal{B}(\mathbf{x}_1, r_1)$ is approximately equal to $f(\mathbf{x}_1)$, which means that $F(r_1|\mathbf{x}_1)$ by Definition 3.1.2 and (3.1) can be approximated as

$$F(r_1|\mathbf{x}_1) = \int_{\|\mathbf{x}_2-\mathbf{x}_1\|<r_1} f(\mathbf{x}_2)\,\mathrm{d}\mathbf{x}_2 \approx f(\mathbf{x}_1) \int_{\|\mathbf{x}_2-\mathbf{x}_1\|<r_1} \mathrm{d}\mathbf{x}_2 = f(\mathbf{x}_1) \cdot V_{r_1}. \tag{3.2}$$

Note that r_1 is actually a random variable since it depends on the kNN of $\mathbf{x}_1$, and therefore $F(r_1|\mathbf{x}_1)$ is also a random variable. Note also that $F(r_1|\mathbf{x}_1)$ can be considered as a mass, and in (3.2) it is approximated as the density multiplied by the volume, which means that the density $f(\mathbf{x}_1)$ is assumed to be constant inside $\mathcal{B}(\mathbf{x}_1, r_1)$. The result in (3.2) leads to

$$\begin{aligned} f(\mathbf{x}_1) \cdot V_{r_1} &\approx P(\mathbf{X}_2 \in \mathcal{B}(\mathbf{x}_1, r_1) \,|\, \mathbf{X}_1 = \mathbf{x}_1) \\ &\approx \frac{1}{n-1} \sum_{i=2}^{n} \mathbb{I}(\mathbf{x}_i \in \mathcal{B}(\mathbf{x}_1, r_1) \,|\, \mathbf{X}_1 = \mathbf{x}_1) = \frac{k-1}{n-1} \Rightarrow \\ f(\mathbf{x}_1) \cdot V_{r_1} &\approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\mathbf{x}_i \in \mathcal{B}(\mathbf{x}_1, r_1) \,|\, \mathbf{X}_1 = \mathbf{x}_1) = \frac{k}{n} \Rightarrow \\ f(\mathbf{x}_1) &\approx \frac{k}{n} \frac{1}{V_{r_1}} = \frac{k\Gamma(d/2+1)}{n\pi^{d/2} r_1^d} \Rightarrow f(\mathbf{x}_i) \approx \frac{k\Gamma(d/2+1)}{n\pi^{d/2} r_i^d}, \end{aligned} \tag{3.3}$$

where the definition of the edf (see (A.2)) has been used for the approximation in the second line and it has been assumed that n is large between the second and third line. By using the definition of the expected value of a random variable in (A.3), the entropy as defined in Definition 2.1.2 can be estimated with the following basic estimator denoted by $\hat{h}_{\mathrm{B}}(\mathbf{X})$, where the approximation of $f(\mathbf{x}_i)$ in (3.3) is inserted

$$\begin{aligned} h(\mathbf{X}) &= -\int_{\mathbb{R}^d} f(\mathbf{x}) \ln(f(\mathbf{x})) \, \mathrm{d}\mathbf{x} = -\mathbb{E}[\ln(f)] \approx -\frac{1}{n} \sum_{i=1}^{n} \ln\left(f(\mathbf{x}_i)\right) \Rightarrow \\ \hat{h}_{\mathrm{B}}(\mathbf{X}) &\triangleq -\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{k\Gamma(d/2+1)}{n\pi^{d/2} r_i^d}\right) = \frac{1}{n} \sum_{i=1}^{n} T_i, \quad T_i \triangleq \ln\left(\frac{n\pi^{d/2} r_i^d}{k\Gamma(d/2+1)}\right). \end{aligned} \tag{3.4}$$

Note that T_i only depends on i through the radius r_i. In Section 3.2, the asymptotic mean of $\hat{h}_{\mathrm{B}}(\mathbf{X})$ is considered.

3.2 An asymptotically unbiased estimator

By finding the asymptotic mean of $\hat{h}_{\mathrm{B}}(\mathbf{X})$, it can be shown that the estimator is asymptotically biased, and an asymptotically unbiased estimator can furthermore be found.

The asymptotic mean of $\hat{h}_{\mathrm{B}}(\mathbf{X})$ can be written as

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}\left[\hat{h}_{\mathrm{B}}(\mathbf{X})\right] &= \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} T_i\right] \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[T_i\right] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[T_i \,|\, \mathbf{X}_i = \mathbf{x}_i] \end{aligned} \tag{3.5}$$

since

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[T_i\,|\,\mathbf{X}_i=\mathbf{x}_i]=\mathbb{E}[\,\mathbb{E}[T_i\,|\,\mathbf{X}_i=\mathbf{x}_i]\,]=\mathbb{E}[T_i]=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,[T_i]\,,$$

where in the first and last equalities the law of large numbers has been used and the law of total expectation has been used in the second equality [14, pp. 190, 264].

Since the random variables $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are assumed to be iid (as described in Section 3.1), the case that $\mathbf{x}_1$ is given in (3.5) is considered without loss of generality in the following. Note that T_1 being larger than a real number t is equivalent to the radius r_1 being larger than a real number ρ_t:

$$T_1=\ln\left(\frac{n\pi^{d/2}r_1^d}{k\Gamma(d/2+1)}\right)>t\Leftrightarrow\frac{n\pi^{d/2}r_1^d}{k\Gamma(d/2+1)}>\exp(t)\Leftrightarrow$$
$$r_1^d>\frac{k\Gamma(d/2+1)\exp(t)}{n\pi^{d/2}}\Leftrightarrow r_1>\left(\frac{k\Gamma(d/2+1)\exp(t)}{n\pi^{d/2}}\right)^{\frac{1}{d}}\triangleq\rho_t.$$

By Definition 3.1.2, $F(\rho_t|\mathbf{x}_1)$ is the conditional probability that a random variable is inside the ball $\mathcal{B}(\mathbf{x}_1,\rho_t)$, where $\mathbf{x}_1$ is given. Consider now the conditional probability that r_1 is larger than ρ_t, where $\mathbf{x}_1$ is given. There are $k-1$ samples inside $\mathcal{B}(\mathbf{x}_1,r_1)$ beside $\mathbf{x}_1$, and these samples can either be inside (with probability $F(\rho_t|\mathbf{x}_1)$) or outside (with probability $1-F(\rho_t|\mathbf{x}_1)$) the ball $\mathcal{B}(\mathbf{x}_1,\rho_t)$. The conditional probability that r_1 is larger than ρ_t, where $\mathbf{x}_1$ is given, can then be written as a sum of k binomial terms (see Definition A.2.2), where in each term, an increasing number of samples up to $k-1$ are chosen out of the remaining $n-1$ samples:

$$\begin{aligned}P(T_1>t\mid\mathbf{X}_1=\mathbf{x}_1)&=P(r_1>\rho_t\mid\mathbf{X}_1=\mathbf{x}_1)\\&=\sum_{i=0}^{k-1}\binom{n-1}{i}\left[F(\rho_t|\mathbf{x}_1)\right]^i\left[1-F(\rho_t|\mathbf{x}_1)\right]^{n-1-i}.\end{aligned}\tag{3.6}$$

By (3.1), the volume of a ball with radius ρ_t can be written as

$$V_{\rho_t}=\frac{\pi^{d/2}\rho_t^d}{\Gamma(d/2+1)}=\frac{\pi^{d/2}\left(\left(\frac{k\Gamma(d/2+1)\exp(t)}{n\pi^{d/2}}\right)^{\frac{1}{d}}\right)^d}{\Gamma(d/2+1)}=\frac{k\exp(t)}{n}.$$

The mean of a binomial distribution with parameters n,p (see Definition A.2.2) is np, which means that the asymptotic mean of each term in the sum in the expression in (3.6) is

$$\lim_{n\to\infty}((n-1)F(\rho_t|\mathbf{x}_1))=\lim_{n\to\infty}(nF(\rho_t|\mathbf{x}_1))$$

$$= k\exp(t)\lim_{n\to\infty}\left(\frac{F(\rho_t|\mathbf{x}_1)}{V_{\rho_t}}\right) = k\exp(t)f(\mathbf{x}_1), \tag{3.7}$$

where $\lim_{n\to\infty}\left(\frac{F(\rho_t|\mathbf{x}_1)}{V_{\rho_t}}\right) = f(\mathbf{x_1})$ because the density can be written as the probability mass divided by the volume for which $V_{\rho_t} \to 0$ as $n \to \infty$.

By using the Poisson approximation to the binomial distribution (see Proposition A.3.1), each of the terms in the sum in (3.6) can be written as the pmf of a Poisson distribution (see Definition A.2.1) with parameter equal to the asymtotic mean $k\exp(t)f(\mathbf{x}_1)$ from (3.7):

$$\begin{aligned}\lim_{n\to\infty} P(T_1 > t \mid \mathbf{X}_1 = \mathbf{x}_1) &= \sum_{i=0}^{k-1}\frac{\left(k\exp(t)f(\mathbf{x}_1)\right)^i}{i!}\exp\left(-k\exp(t)f(\mathbf{x}_1)\right)\\ &= P(T_{\mathbf{x}_1} > t),\end{aligned}$$

where $T_{\mathbf{x}_1}$ is a random variable with the pdf $g(y)$

$$g(y) = \frac{\left(k\exp(y)f(\mathbf{x}_1)\right)^k}{(k-1)!}\exp(-k\exp(y)f(\mathbf{x}_1)), \quad -\infty < y < \infty. \tag{3.8}$$

In order to verify that this is the correct pdf, it must be shown that

$$\begin{aligned}P(T_{\mathbf{x}_1} > t) &= \int_t^\infty \frac{\left(k\exp(y)f(\mathbf{x}_1)\right)^k}{(k-1)!}\exp(-k\exp(y)f(\mathbf{x}_1))\,\mathrm{d}y\\ &= \sum_{i=0}^{k-1}\frac{\left(k\exp(t)f(\mathbf{x}_1)\right)^i}{i!}\exp\left(-k\exp(t)f(\mathbf{x}_1)\right).\end{aligned} \tag{3.9}$$

The first equality is simply the integral of the proposed pdf, and the second equality can be shown by using integration by substitution and integration by parts (see Lemmas A.4.1 and A.4.2). Integration by substitution is firstly used to obtain the integral

$$\begin{aligned}P(T_{\mathbf{x}_1} > t) &= \int_t^\infty \frac{\left(k\exp(y)f(\mathbf{x}_1)\right)^k}{(k-1)!}\exp(-k\exp(y)f(\mathbf{x}_1))\,\mathrm{d}y\\ &= \int_{z(t)}^\infty \frac{(z(y))^{k-1}}{(k-1)!}\exp(-z(y))\,\mathrm{d}z(y),\end{aligned} \tag{3.10}$$

$$z(y) = k\exp(y)f(\mathbf{x}_1) \Rightarrow \frac{\mathrm{d}}{\mathrm{d}y}z(y) = z(y) \Rightarrow \mathrm{d}y = \frac{1}{z(y)}\,\mathrm{d}z(y),$$

where the integral boundaries in the new integral are obtained by evaluating $z(y)$ with y equal to each of the integral boundaries in the original integral.

As shown in Appendix A.4, repeated integration by parts k times can be written as (see (A.5))

$$\int_{z(t)}^\infty u^{(0)}(z(y))\; v^{(k)}(z(y))\,\mathrm{d}z(y) = \sum_{i=0}^{k-1}(-1)^i\left(u^{(i)}(\infty)\; v^{(k-1-i)}(\infty)\right.$$

$$\left. -u^{(i)}(z(t))\ v^{(k-1-i)}(z(t)) \right)$$
$$+ (-1)^k \int_{z(t)}^{\infty} u^{(k)}(z(y))\ v^{(0)}(z(y))\,\mathrm{d}z(y), \qquad (3.11)$$

where $u^{(i)}(z(t))$ denotes the i'th derivative of $u(z(t))$.

By comparing (3.10) to (3.11), we let $u(z(y)) = \frac{(z(y))^{k-1}}{(k-1)!}$ and $v^{(k)}(z(y)) = \exp(-z(y))$. The k derivatives of $u(z(y))$ and k antiderivatives of $v^{(k)}(z(y))$ are then needed in order to perform integration by parts k times.

The first three derivatives of $u(z(y))$ are

$$u(z(y)) = \frac{(z(y))^{k-1}}{(k-1)!} \Rightarrow u^{(1)}(z(y)) = \frac{(k-1)(z(y))^{k-2}}{(k-1)!} = \frac{(z(y))^{k-2}}{(k-2)!}$$
$$u^{(2)}(z(y)) = \frac{(k-1)(k-2)(z(y))^{k-3}}{(k-1)!} = \frac{(z(y))^{k-3}}{(k-3)!}$$
$$u^{(3)}(z(y)) = \frac{(k-1)(k-2)(k-3)(z(y))^{k-4}}{(k-1)!} = \frac{(z(y))^{k-4}}{(k-4)!},$$

which generally means that

$$u^{(i)}(z(y)) = \frac{(z(y))^{k-1-i}}{(k-1-i)!}, \quad i = 0, \ldots, k-1.$$

Note that $u^{(k-1)}(z(y)) = 1$, which means that $u^{(k)}(z(y)) = 0$.

The first three antiderivatives of $v^{(k)}(z(y))$ are

$$v^{(k)}(z(y)) = \exp(-z(y)) \Rightarrow v^{(k-1)}(z(y)) = -\exp(-z(y))$$
$$v^{(k-2)}(z(y)) = \exp(-z(y))$$
$$v^{(k-3)}(z(y)) = -\exp(-z(y)),$$

which generally means that

$$v^{(k-2i)}(z(y)) = \exp(-z(y)), \quad v^{(k-(2i+1))}(z(y)) = -\exp(-z(y)), \quad i = 0, 1, \ldots, \left\lfloor \frac{k}{2} \right\rfloor.$$

With regards to (3.11), note that

1. $(-1)^i \cdot v^{(k-1-i)}(z(y)) = -\exp(-z(y))$ for all $i = 0, \ldots, k-1$ due to the derivation of the antiderivatives of $v^{(k)}(z(y))$.
2. Since $u^{(i)}(\infty)\ v^{(k-1-i)}(\infty) = 0$[1], the first part of (3.11) vanishes.

[1] This can be verified from plots of the expression.

3. Since $u^{(k)}(z(y)) = 0$, the integral in the last part of (3.11) vanishes.

With these considerations and the results above, the desired probability $P(T_{\mathbf{x}_1} > t)$ can finally be written as

$$\begin{aligned}
P(T_{\mathbf{x}_1} > t) &= \int_{z(t)}^{\infty} \frac{(z(y))^{k-1}}{(k-1)!} \exp(-z(y))\, \mathrm{d}z(y) \\
&= \int_{z(t)}^{\infty} u^{(0)}(z(y))\ v^{(k)}(z(y))\, \mathrm{d}z(y) \\
&= \sum_{i=0}^{k-1} (-1)^i \left(-u^{(i)}(z(t))\ v^{(k-1-i)}(z(t)) \right) \\
&= \sum_{i=0}^{k-1} \frac{(z(t))^i}{i!} \exp(-z(t)) \\
&= \sum_{i=0}^{k-1} \frac{(k \exp(t) f(\mathbf{x}_1))^i}{i!} \exp\left(-k \exp(t) f(\mathbf{x}_1)\right),
\end{aligned}$$

which is the desired expression from (3.9). This means that the pdf $g(y)$ in (3.8) is in fact the right one.

By using integration by substitution and (A.4), the asymptotic mean of T_1 given $\mathbf{x}_1$ is

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[T_1 \mid \mathbf{X}_1 = \mathbf{x}_1\right] &= \int_{-\infty}^{\infty} y \frac{(k \exp(y) f(\mathbf{x}_1))^k}{(k-1)!} \exp(-k \exp(y) f(\mathbf{x}_1))\, \mathrm{d}y \\
&= \int_0^{\infty} (\ln(z) - \ln(k) - \ln(f(\mathbf{x}_1)) \frac{z^{k-1}}{(k-1)!} \exp(-z)\, \mathrm{d}z \\
&= \frac{1}{\Gamma(k)} \int_0^{\infty} \ln(z) z^{k-1} \exp(-z)\, \mathrm{d}z \\
&\quad - (\ln(k) + \ln(f(\mathbf{x}_1)) \frac{1}{(k-1)!} \int_0^{\infty} z^{k-1} \exp(-z)\, \mathrm{d}z \\
&= \psi(k) - \ln(k) - \ln(f(\mathbf{x}_1)), \qquad (3.12)
\end{aligned}$$

where the change of variable $z = k \exp(y) f(\mathbf{x}_1)$[2] has been made, which means that $y = \ln(z) - \ln(k) - \ln(f(\mathbf{x}_1))$, which in turn means that $\mathrm{d}y = \frac{1}{z}\, \mathrm{d}z$, and where

$$\psi(\nu) = \frac{\Gamma'(\nu)}{\Gamma(\nu)}, \qquad \Gamma'(\nu) = \int_0^{\infty} \ln(t) t^{\nu-1} \exp(-t)\, \mathrm{d}t,$$
$$\Gamma(\nu) = \int_0^{\infty} t^{\nu-1} \exp(-t)\, \mathrm{d}t = (\nu-1)!$$

[2]Note that $0 \leq z < \infty$ since $k \geq 1$, $f(\mathbf{x}_1) \geq 0$, and $\exp(y) > 0$.

is the digamma function, which is defined as the derivative of the gamma function divided by the gamma function [12, eqs. 5.2.1, 5.2.2, 5.4.1, 5.9.19].

The results in (3.5) and (3.12) lead to that the asymptotic mean of $\hat{h}_{\mathrm{B}}(\mathbf{X})$ is

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\hat{h}_{\mathrm{B}}(\mathbf{X})\right] &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[T_i \mid \mathbf{X}_i = \mathbf{x}_i] \\ &= \lim_{n\to\infty} \psi(k) - \ln(k) - \frac{1}{n}\sum_{i=1}^{n} \ln(f(\mathbf{x}_i)) \\ &= \psi(k) - \ln(k) + h(\mathbf{X}).\end{aligned}$$

In Section 3.4 it is shown that $\lim_{k\to\infty} \psi(k) - \ln(k) = 0$, which means that the estimator $\hat{h}_{\mathrm{B}}$ is in fact unbiased for large values of k since asymptotic mean is then equal to the true value, $h(\mathbf{X})$. However, k is assumed to be small for now.

The difference between the asymptotic mean and $h(\mathbf{X})$ is $\psi(k) - \ln(k)$, and an asymptotically unbiased estimator can then be obtained (as proposed in [18, p. 307]) by subtracting this from $\hat{h}_{\mathrm{B}}(\mathbf{X})$, which is denoted as the estimator $\hat{h}_{\mathrm{S}}(\mathbf{X})$

$$\begin{aligned}\hat{h}_{\mathrm{S}}(\mathbf{X}) &= \hat{h}_{\mathrm{B}}(\mathbf{X}) - (\psi(k) - \ln(k)) \\ &= \frac{1}{n}\sum_{i=1}^{n} T_i - \psi(k) + \ln(k) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{n\pi^{d/2} r_i^d}{k\Gamma(d/2+1)}\right) - \psi(k) + \ln(k) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(r_i^d\right) + \ln\left(\frac{\pi^{d/2}}{\Gamma(d/2+1)}\right) + \ln(n) - \ln(k) - \psi(k) + \ln(k) \\ &= \frac{d}{n}\sum_{i=1}^{n} \ln(r_i) + \ln(V_1) + \ln(n) - \psi(k), \end{aligned} \tag{3.13}$$

where V_1 is the volume of the d-dimensional unit ball.

It is furthermore proven in [18, p. 307] that $\hat{h}_{\mathrm{S}}(\mathbf{X})$ is also asymptotically consistent, which means that

$$\lim_{n\to\infty} \mathrm{Var}\left[\hat{h}_{\mathrm{S}}(\mathbf{X})\right] = 0.$$

Being asymptotically unbiased and consistent is obviously good qualities for an estimator.

3.3 General derivation of estimators

This section describes a different approach to deriving an entropy estimator than the one described in Section 3.2. The approach described in this section is inspired by

[11, p. 2] [9, pp. 3-4][3] and serves as a foundation for the derivations of the estimators described in Section 3.4, where a constant density is assumed. The general setup described in Section 3.1 is also used in the remainder of this chapter.

The setup shown on Figure 3.1 is firstly considered. The conditional probability given $\mathbf{x}_1$ that only the kNN lies in the annulus $[r_1, r_1 + \mathrm{d}r_1]$ around $\mathcal{B}(\mathbf{x}_1, r_1)$, that exactly $k-1$ samples are at distances closer to $\mathbf{x}_1$ than the kNN, and that the remaining $n-k-1$ samples are farther away than the kNN, is denoted by $f(r_1|\mathbf{x}_1)\mathrm{d}r_1$, which consists of a conditional density $f(r_1|\mathbf{x}_1)$[4] multiplied by the width $\mathrm{d}r_1$ of the annulus shown on Figure 3.1. This conditional probability given $\mathbf{x}_1$ can then be expressed as [11, p. 2]

$$f(r_1|\mathbf{x}_1)\mathrm{d}r_1 = \binom{n-1}{1} \frac{\mathrm{d}F(r_1|\mathbf{x}_1)}{\mathrm{d}r_1} \mathrm{d}r_1 \binom{n-2}{k-1} [F(r_1|\mathbf{x}_1)]^{k-1} \cdot [1 - F(r_1|\mathbf{x}_1)]^{n-k-1} \tag{3.14}$$

where the first binomial coefficient in (3.14) expresses in how many ways the kNN can be chosen from the $n-1$ samples (where $\mathbf{x}_1$ is given) and the second binomial coefficient expresses in how many ways the $k-1$ samples (except $\mathbf{x}_1$) inside the ball $\mathcal{B}(\mathbf{x}_1, r_1)$ can be chosen from the remaining $n-2$ samples. Furthermore, the last part in (3.14) is due to the samples being iid, where $[F(r_1|\mathbf{x}_1)]^{k-1}$ and $[1-F(r_1|\mathbf{x}_1)]^{n-k-1}$ are the probabilities that $k-1$ and $n-k-1$ samples lie inside and outside $\mathcal{B}(\mathbf{x}_1, r_1)$, respectively. The expression for $f(r_1|\mathbf{x}_1)\mathrm{d}r_1$ in (3.14) is however not used in the following but mainly serves as a way of understanding what $f(r_1|\mathbf{x}_1)\mathrm{d}r_1$ is.

In the following, the definition of the conditional expected value (see (A.4)) is used to express the expected value of $\ln(F(r_1|\mathbf{x}_1))$ where $\mathbf{x}_1$ is given

$$\mathbb{E}[\ln(F(r_1|\mathbf{x}_1)) \mid \mathbf{X}_1 = \mathbf{x}_1] = \int_0^\infty f(r_1|\mathbf{x}_1) \ln(F(r_1|\mathbf{x}_1)) \, \mathrm{d}r_1 = \psi(k) - \psi(n). \tag{3.15}$$

It has however not been possible to verify the result in (3.15) analytically. In [9, p. 4] and [11, p. 2], expressions similar to the one in (3.14) is used but it is not further described how to obtain the result. In this thesis, an approach for computing the integral in (3.15) has instead been devised. This is described in Section 3.3.1, and in the remainder of this section the result in (3.15) is assumed to be true since it makes it possible to derive another estimator.

As shown in (3.2), $F(r_1|\mathbf{x}_1)$ can be approximated as the density multiplied by the volume of $\mathcal{B}(\mathbf{x}_1, r_1)$ if the density is assumed to be constant. With no assumptions

[3] The notation in the cited articles is however not clear, and the notation in this chapter is made by the author.

[4] This means that $\mathbf{X}_1 = \mathbf{x}_1$ is given but is written as $f(r_1|\mathbf{x}_1)$ for ease of notation (similarly to Definition 3.1.2).

on the density, $F(r_1|\mathbf{x}_1)$ can be approximated by a random variable η_1 multiplied by the density

$$F(r_1|\mathbf{x}_1) \approx \eta_1 f(\mathbf{x}_1). \tag{3.16}$$

Assuming that the result in (3.15) is true, it follows that

$$\begin{aligned}
\mathbb{E}[\ln(F(r_1|\mathbf{x}_1)) \,|\, \mathbf{X}_1 = \mathbf{x}_1] &= \psi(k) - \psi(n) \\
&\approx \mathbb{E}[\ln(\eta_1 f(\mathbf{x}_1))] \\
&= \mathbb{E}[\ln(\eta_1) + \ln(f(\mathbf{x}_1))] \\
&= \mathbb{E}[\ln(\eta_1)] + \mathbb{E}[\ln(f(\mathbf{x}_1))] \Rightarrow \\
-\mathbb{E}[\ln(f(\mathbf{x}_1))] &\approx \mathbb{E}[\ln(\eta_1)] + \psi(n) - \psi(k) \Rightarrow \\
h(\mathbf{X}) &= \lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^{n} \ln(f(\mathbf{x}_i)) \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \ln(\eta_i) + \psi(n) - \psi(k),
\end{aligned} \tag{3.17}$$

where it has been used that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for any two random variables X, Y [14, p. 176], and where η_i, $i = 1, \ldots, n$, depends on which assumptions of the density $f(\mathbf{x}_i)$ inside $\mathcal{B}(\mathbf{x}_i, r_i)$ are made.

3.3.1 Numerical computation of integral

The procedure for computing the integral in (3.15) numerically is described in this section.

In general, the functions $f(r_1|\mathbf{x}_1)$ and $F(r_1|\mathbf{x}_1)$ are estimated as $\hat{f}(r_1|\mathbf{x}_1)$ and $\hat{F}(r_1|\mathbf{x}_1)$ from n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of iid random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X}_i \in \mathbb{R}^d$ and with a given distribution. The idea is to estimate $\hat{f}(r_1|\mathbf{x}_1)$ and $\hat{F}(r_1|\mathbf{x}_1)$ with different values of r_1 in some appropriate interval corresponding to the minimum and maximum distances between $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The density $f(r_1|\mathbf{x}_1)$ is then estimated as the histogram of r_1 (which is $\hat{f}(r_1|\mathbf{x}_1)$ and which includes finding the frequency of r_1 in each bin), while $\hat{F}(r_1|\mathbf{x}_1)$ is estimated as the ratio of samples within distance $r_1 > 0$ from $\mathbf{x}_1$ (due to Definition 3.1.2).

Note that $f(r_1|\mathbf{x}_1)\,\mathrm{d}r_1$ by (3.14) depends on n and k. As described in Section 3.1, k is the number of samples inside the ball $\mathcal{B}(\mathbf{x}_1, r_1)$, and k is therefore changed when r_1 is changed, which means that the integral will not depend on k (contrary to the statement in (3.15)). The expression in (3.14) also implicitly depends on d since $\mathbf{x}_1$ is given. Therefore, it is proposed to compute the integral in (3.15) numerically where either n or d is changed, while the other is fixed. By doing so for d, n in a specified range of values, a grid $\mathbf{M}$ can be obtained, where $\mathbf{M}(d, n)$ is the integral for the specified values of d, n.

The procedure for computing the integral is described in Algorithm 1. Note that objects such as a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ are indexed with row r and column c as $\mathbf{A}(r, c)$ and that $\mathbf{A}(:, c)$ and $\mathbf{A}(r, :)$ returns the entire column c and row r, respectively.
In general, the values

$$n \in \{25, 50, 75, 100, 1100, \ldots, 8{,}100, 9{,}100\}, \qquad d \in \{1, 5, \ldots, 41, 45\}$$

are used in Algorithm 1, and examples of $\hat{f}(r_1|\mathbf{x}_1)$ and $\hat{F}(r_1|\mathbf{x}_1)$ are shown in the following for a fixed value of n and different values of d.

Figure 3.2 shows $\hat{f}(r_1|\mathbf{x}_1)$ as the histogram of r_1 for $n = 1{,}100$ random variables in different dimensions d and with a uniform distribution (left) and a multivariate normal distribution (right).

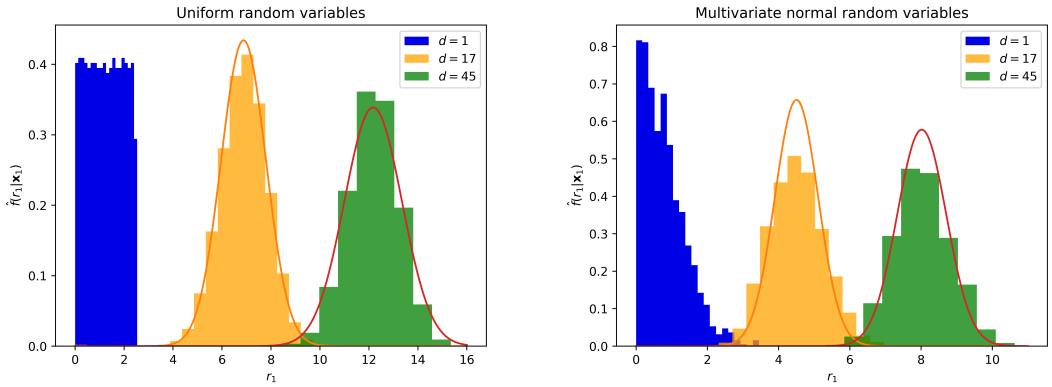


Figure 3.2: Histograms of r_1 for different values of d with $n = 1{,}100$ random variables with a uniform distribution (left) and multivariate normal distribution. Curves for the normal density with mean and variance corresponding to each histogram is also shown for the histograms with $d > 1$.

The results in Figure 3.2 are generally as expected since the distances between samples of random variables with a multivariate normal distribution is also a normal distribution. Furthermore, the histogram for the distances with a uniform distribution with $d = 1$ is constant but for $d > 1$, these distances also have a normal distribution.

Figure 3.3 shows $\hat{F}(r_1|\mathbf{x}_1)$ for $n = 1{,}100$ random variables in different dimensions d and with a uniform distribution (left) and a multivariate normal distribution (right).
The results in Figure 3.2 are also as expected since the cdf for a normal distribution,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \mathrm{d}x,$$

is a logistic function for which $\lim_{x \to -\infty} \Phi(x) = 0$ and $\lim_{x \to \infty} \Phi(x) = 1$ (similarly to the results in Figure 3.3) [15].

Note that the functions are only computed for r_1 in an interval where the functions are different from 0 and 1 (as described in Algorithm 1), which is the reason

Algorithm 1 Algorithm for numerical computation of integral

Input: the ranges of d-values $\mathbf{d}$ and n-values $\mathbf{n}$ to be tested, the distribution of the random variables, the number of test iterations N, the minimum and maximum distances m_1, m_2, and the number of bins b.
Let l_1, l_2 be the lengths of $\mathbf{d}, \mathbf{n}$ and let

$$\mathbf{M} \in \mathbb{R}^{l_1 \times l_2}, \quad \mathbf{r}, \mathbf{f}_m, \mathbf{F}_m \in \mathbb{R}^{l_1 \times l_2 \times l_2}, \quad \mathbf{f}, \mathbf{f}_i, \mathbf{F}, \mathbf{F}_l \in \mathbb{R}^{l_1 \times l_2 \times N \times l_2}, \quad \mathbf{f}_h \in \mathbb{R}^{l_1 \times l_2 \times b}$$

be filled with zeros.
for each d in $\mathbf{d}$, n in $\mathbf{n}$ **do**
  Fill $\mathbf{r}(d, n, :) \in \mathbb{R}^n$ with equidistant values in the interval $[m_1, m_2]$.
  for $t = 1, \ldots, N$ **do**
    Find samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the n random variables in dimension d.
    Define $\mathbf{x}_1 \in \mathbb{R}^d$ as the center sample of the samples.
    Create the matrix $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\mathsf{T} \in \mathbb{R}^{n \times d}$.
    $\mathbf{D} \leftarrow \text{sort}(\|\mathbf{x} - \mathbf{x}_1\|_2)$ (i.e. the sorted distances).
    $\mathbf{f}(d, n, t, :) \leftarrow \mathbf{D}$.
    for each r in $\mathbf{r}$ **do**
      Let k be the number of samples within distance r from $\mathbf{x}_1$.
      $\mathbf{F}(d, n, t, r) \leftarrow k/n$.
for each d in $\mathbf{d}$, n in $\mathbf{n}$, r in $\mathbf{r}$ **do**

$$\mathbf{F}_m(d, n, r) \leftarrow \frac{1}{N} \sum_{t=1}^{N} \mathbf{F}(d, n, t, r)$$

$$\mathbf{f}_m(d, n, r) \leftarrow \frac{1}{N} \sum_{t=1}^{N} \mathbf{f}(d, n, t, r)$$

for each d in $\mathbf{d}$, n in $\mathbf{n}$ **do**
  $\mathbf{F}_l(d, n, :) \leftarrow \ln(\mathbf{F}_m(d, n, :))$ (with $\ln(0) = 0$).
  Set $\mathbf{f}_h(d, n, :)$ to the frequencies of a histogram of $\mathbf{f}_m(d, n, :)$ with b bins.
  Let $\mathbf{f}_i(d, n, :)$ be an interpolated version of $\mathbf{f}_h(d, n, :)$.
  Compute $\mathbf{M}(d, n)$ through numerical integration of $\mathbf{f}_i(d, n, :) \cdot \mathbf{F}_l(d, n, :)$.
Output: the matrix $\mathbf{M}$.

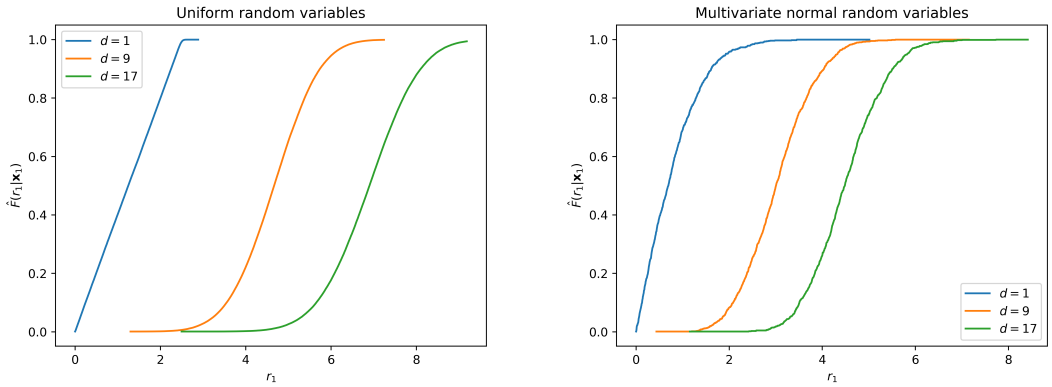


Figure 3.3: Plots of $\hat{F}(r_1|\mathbf{x}_1)$ for different values of d with $n = 1{,}100$ random variables with a uniform distribution (left) and multivariate normal distribution.

for the shortened lines in Figure 3.3. Since $\hat{f}(r_1|\mathbf{x}_1)$ is 0 outside this interval, the product $\hat{f}(r_1|\mathbf{x}_1)\ln(\hat{F}(r_1|\mathbf{x}_1))$ is also 0 outside this interval, which means that there is no reason to compute the values of $\hat{f}(r_1|\mathbf{x}_1)$ and $\hat{F}(r_1|\mathbf{x}_1)$ outside this interval.

After $\hat{f}(r_1|\mathbf{x}_1)$ has been estimated, its values are interpolated such that the same number of output values are available for $\hat{f}(r_1|\mathbf{x}_1)$ and $\hat{F}(r_1|\mathbf{x}_1)$. Furthermore, $\ln(\hat{F}(r_1|\mathbf{x}_1))$ is found, where $\ln(0)$ is set to 0. The outputs of the functions $\hat{f}(r_1|\mathbf{x}_1)$ and $\ln(\hat{F}(r_1|\mathbf{x}_1))$ are then multiplied to form another function, which is integrated between 0 and the maximum value of its support.

Figure 3.4 shows the mean values and their confidence intervals of the integrals of $\hat{f}(r_1|\mathbf{x}_1)\ln(\hat{F}(r_1|\mathbf{x}_1))$, where the mean is taken across the values of n (left plots) and across the values of d (right plots), and where the random variables are uniformly distributed (top plots) and multivariate normally distributed (bottom plots).
Figure 3.4 indicates that the confidence intervals are relatively close to the mean values regardless of which axis the mean is taken across, which means that it may not be necessary to use the matrix $\mathbf{M}(d, n)$ as the value of the integral for specified values of d, n as suggested in Algorithm 1. Instead, it might only be necessary to take the mean across both n and d and use this value as the integral of $f(r_1|\mathbf{x}_1)\ln(F(r_1|\mathbf{x}_1))$ (no matter what the value of n and d is). Therefore, the constants M_u and M_n are defined as the mean values of the grids with the random variables with a uniform and multivariate normal distribution, respectively. These constants and the corresponding 95% confidence intervals are:

$$M_\text{u} = -0.82, \quad [-0.78, -0.86],$$
$$M_\text{n} = -0.82, \quad [-0.79, -0.85].$$

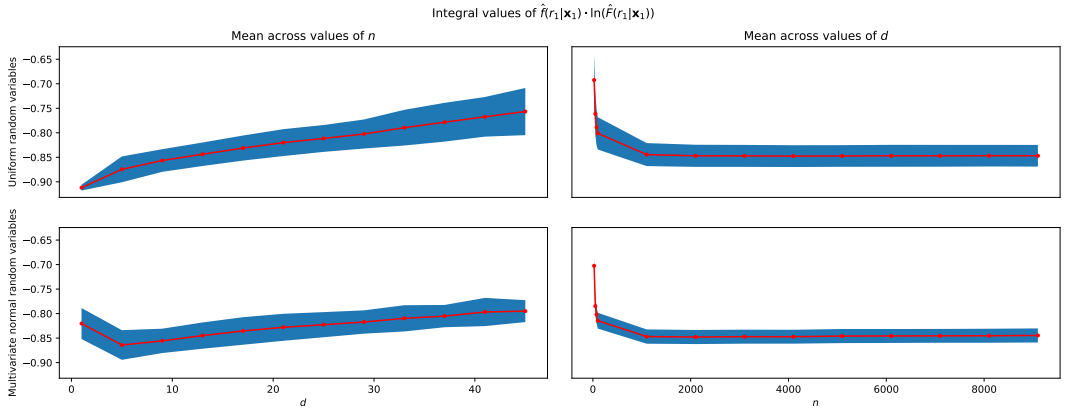


Figure 3.4: Mean values and the corresponding confidence intervals of the integrals of $\hat{f}(r_1|\mathbf{x}_1) \cdot \ln(\hat{F}(r_1|\mathbf{x}_1))$ across both n and d and with both uniform and multivariate normal distributions of the random variables.

3.4 Estimator for constant density

If the probability density $f(\mathbf{x}_1)$ is assumed to be constant inside $\mathcal{B}(\mathbf{x}_1, r_1)$, the conditional probability $F(r_1|\mathbf{x}_1)$ can be approximated as the volume of $\mathcal{B}(\mathbf{x}_1, r_1)$ multiplied by the density (as in (3.2))

$$F(r_1|\mathbf{x}_1) \approx V_{r_1} f(\mathbf{x}_1) = V_1 r_1^d f(\mathbf{x}_1). \tag{3.18}$$

With the approximation in (3.18), the constant η_1 in (3.16) is $\eta_1 = V_1 r_1^d$, which is combined with the result in (3.17) to obtain the entropy estimator $\hat{h}_{\mathrm{KL}}(\mathbf{X})$ [11, p. 2]

$$\begin{aligned}\hat{h}_{\mathrm{KL}}(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(V_1 r_i^d\right) + \psi(n) - \psi(k) \\ &= \frac{d}{n}\sum_{i=1}^{n} \ln(r_i) + \ln(V_1) + \psi(n) - \psi(k).\end{aligned} \tag{3.19}$$

However, the estimator $\hat{h}_{\mathrm{KL}}$ relies on the derivation in (3.17), which assumes that the result in (3.15) is true. By instead using the constant M_{u}, which is derived from the numerical analysis in Section 3.3.1 and computed with uniformly distributed random variables, another estimator denoted by $\hat{h}_{\mathrm{Mu}}$ can be derived as

$$\hat{h}_{\mathrm{Mu}}(\mathbf{X}) = \frac{d}{n}\sum_{i=1}^{n} \ln(r_i) + \ln(V_1) + M_{\mathrm{u}}. \tag{3.20}$$

Note that the only difference between the two estimators $\hat{h}_{\mathrm{KL}}$ and $\hat{h}_{\mathrm{Mu}}$ is the function $\psi(n) - \psi(k)$ and the constant M_{u}, respectively. The function $\psi(n) - \psi(k)$ will always

be positive since $k \leq n-1$ is a positive integer and $\psi(n)$ is positive and monotonically increasing for all $n \in \mathbb{N}$ (see Figure 3.5), whereas M_{u} is a negative constant.

The name of the estimator $\hat{h}_{\mathrm{KL}}$ is due to its original inventors, Kozachenko and Leonenko [11, p. 2]. Note that the only difference between the estimators $\hat{h}_{\mathrm{S}}$ and $\hat{h}_{\mathrm{KL}}$ from (3.13) and (3.19), respectively, is the terms $\ln(n)$ and $\psi(n)$, respectively. However, the two functions

$$\ln(n) - \frac{1}{2n} - \psi(n), \qquad \frac{1}{n} - \ln(n) + \psi(n), \qquad n \in \mathbb{N}$$

are both completely monotonic, which means that [23]

$$\ln(n) - \frac{1}{n} \leq \psi(n) \leq \ln(n) - \frac{1}{2n}.$$

Therefore, $\ln(n) \to \psi(n)$ when n is increased. Figure 3.5 shows the functions $\ln(n)$ and $\psi(n)$ (left) and the absolute difference between them (right).

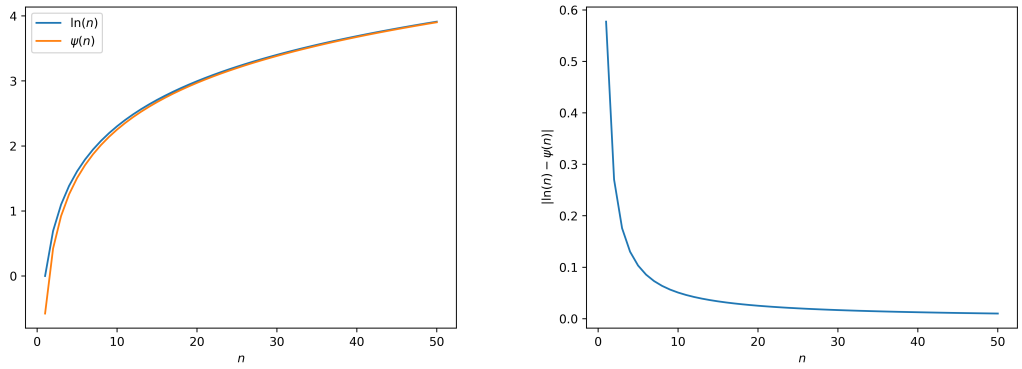


Figure 3.5: Plots of $\ln(n)$ and $\psi(n)$ (left) and the absolute difference between them (right).

As shown in Figure 3.5, the absolute difference between $\ln(n)$ and $\psi(n)$ is rapidly decreasing for small values of n and is asymptotically going to 0. Since $n \in \mathbb{N}$, the largest difference is $|\ln(1) - \psi(1)| = 0.58$. Therefore, the estimators $\hat{h}_{\mathrm{S}}$ and $\hat{h}_{\mathrm{KL}}$ are actually asymptotically the same (but they are derived with different methods, which is why they are both included in this thesis), and only the estimator $\hat{h}_{\mathrm{KL}}$ is used in the analyses in Chapter 4 because the values n in this chapter are rather large (up to 5,100).

4 | Performance analysis

In this chapter the performance of the estimators described in Chapter 3 are analyzed. In Section 4.1, the errors and the computation times of the estimators are analyzed, and computations of the causal conditional directed information (see Section 2.2) are then computed for synthetic autoregressive data and actual EEG data in Sections 4.2 and 4.3, respectively.

4.1 Estimation of entropies

In this section, the entropies of random variables with multivariate normal and uniform distributions are estimated in order to assess the estimators' ability to compute reasonable estimates. Furthermore, the influence of the parameters d, k, and n on the errors and the computation times are analyzed in Sections 4.1.1 and 4.1.2, respectively.

Unless otherwise stated, each of the estimates are computed $N = 200$ times with the parameters $d = 4$, $k = 4$, and $n = 1{,}000$, and in each analysis of these three parameters, two of them are fixed while the third is changed.

4.1.1 Analyses of errors

The error being analyzed in this section is the mean absolute error (MAE)

$$e_{\mathrm{MAE}}\left(h(\mathbf{X}), \hat{h}(\mathbf{X})\right) = \frac{1}{N}\sum_{i=1}^{N}\left|h(\mathbf{X}) - \hat{h}_i(\mathbf{X})\right|,$$

where $\hat{h}_i(\mathbf{X})$ is an estimate of the entropy by one of the estimators described in Chapter 3 and the actual entropy $h(\mathbf{X})$ is obtained from either of the expressions for the entropies of random variables with multivariate normal and uniform distributions in Examples 2.1 and 2.2, respectively, which are refreshed here.

1. Let $\mathbf{X}_{\mathrm{n}}$ be a vector of n mutually independent random variables with a multivariate normal distribution, i.e. $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ (where $\mathbf{I}_n$ is the identity matrix). The analytical value of the entropy of $\mathbf{X}_{\mathrm{n}}$ is (see Example 2.1)

$$\begin{aligned} h(\mathbf{X}_{\mathrm{n}}) &= \frac{1}{2}\ln\left((2\pi\exp(1))^n|\mathbf{\Sigma}|\right) = \frac{1}{2}\ln\left((2\pi\exp(1))^n\right) = \frac{n}{2}\ln((2\pi\exp(1)) \\ &= \frac{n}{2}(\ln(2\pi) + \ln(\exp(1))) = \frac{n}{2}(\ln(2\pi) + 1) = 1.419 \cdot n. \end{aligned}$$

The last expression for $h(\mathbf{X}_\mathrm{n})$ is useful for large values of n since it is not possible to compute $(2\pi \exp(1))^n$ directly for n larger than around 250.

2. Let $R_{\mathbf{a},\mathbf{b}}$ be a hyperrectangle in $\mathbb{R}^d$ whose vertices' lowest and highest coordinates in each dimension are defined by vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Also, let $\mathbf{X}_\mathrm{u}$ be a vector of n mutually independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with a uniform distribution on such a hyperrectangle and with $\mathbf{X}_i \in \mathbb{R}^d$. The analytical value of the entropy of $\mathbf{X}_\mathrm{u}$ is then (see Example 2.2)

$$h(\mathbf{X}_\mathrm{u}) = \ln(V(R_{\mathbf{a},\mathbf{b}})), \qquad V(R_{\mathbf{a},\mathbf{b}}) = |a_1 - b_1| \cdot |a_2 - b_2| \cdots |a_d - b_d|.$$

In this section, $\mathbf{a} = \mathbf{0}$ and $\mathbf{b} = \mathbf{5}$ are fixed.

The figures in this section generally show the MAE's as well as the confidence intervals for each of the computed MAE's (see Appendix A.5). The critical values of a t-distribution is used to compute 95% confidence intervals, where the sample mean and variance is computed from the N absolute errors (which correspond to the random variables that the confidence intervals are computed for), which means that the degrees of freedom are $N - 1$. In Appendix A.5, it is assumed that the data being analyzed are normally distributed, which is seen to be a reasonable assumption from results of statistical tests of whether the errors can be assumed to be normally distributed shown in Appendix B. It however turns out that the confidence intervals are not visible because they are so close to MAE's.

Figure 4.1 shows the MAE's and corresponding confidence intervals of the estimates $\hat{h}(\mathbf{X}_\mathrm{n})$ and $\hat{h}(\mathbf{X}_\mathrm{u})$ with values of $d \in \{1, 5, 9, \ldots, 41, 45\}$.
As shown in Figure 4.1, the mean absolute errors of the estimates increase linearly with d for both $\hat{h}(\mathbf{X}_\mathrm{n})$ and $\hat{h}(\mathbf{X}_\mathrm{u})$ (except for $d = 1$ for the estimator $\hat{h}_\mathrm{Mu}$). This linear relationship means that the estimators only compute the same errors in each dimension. The average increases of error per dimension for the estimates of $\hat{h}(\mathbf{X}_\mathrm{n})$ and $\hat{h}(\mathbf{X}_\mathrm{u})$ are 1.51 and 1.91, respectively, and the average coefficients of determination for these linear relationships are 9.99e−02 and 9.93e−02, respectively. Apart from when $d = 1$, the estimator $\hat{h}_\mathrm{Mu}$ is seen to perform best.

Figure 4.2 shows the MAE's and corresponding confidence intervals of the estimates $\hat{h}(\mathbf{X}_\mathrm{n})$ and $\hat{h}(\mathbf{X}_\mathrm{u})$ with values of $k \in \{1, 2, \ldots, 20\}$.
As shown in Figure 4.2, the errors for the estimators $\hat{h}_\mathrm{B}$ and $\hat{h}_\mathrm{KL}$ are similar and does not change considerably when k is changed whereas the errors of the estimator $\hat{h}_\mathrm{Mu}$ are much lower and has a minimum at $k = 3$ for the uniform random variables (MAE of 2.62e−02) and at $k = 10$ for the multivariate normal random variables (MAE of 4.52e−02). Note also that the estimators are the same for large values of k due to (3.13) and the relationship between the functions $\psi(k)$ and $\ln(k)$ as described in Section 3.4.

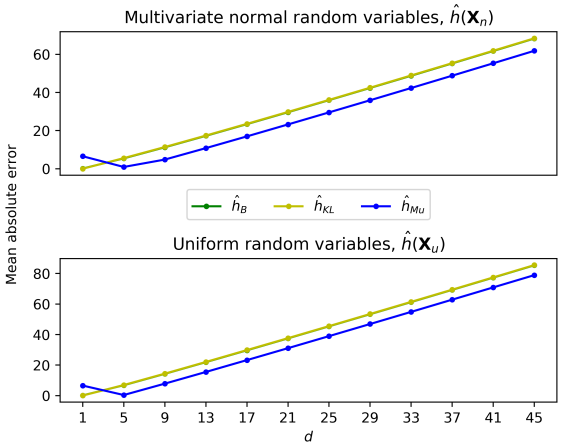


Figure 4.1: The MAE's and corresponding confidence intervals of $N = 200$ entropy estimates of random variables with multivariate normal and uniform distributions and with $d \in \{1, 5, 9, \ldots, 41, 45\}$. The confidence intervals are not directly visible but is e.g. $[68.24, 68.29]$ for the estimator $\hat{h}_{\mathrm{B}}$ with multivariate normal random variables and with $d = 45$.

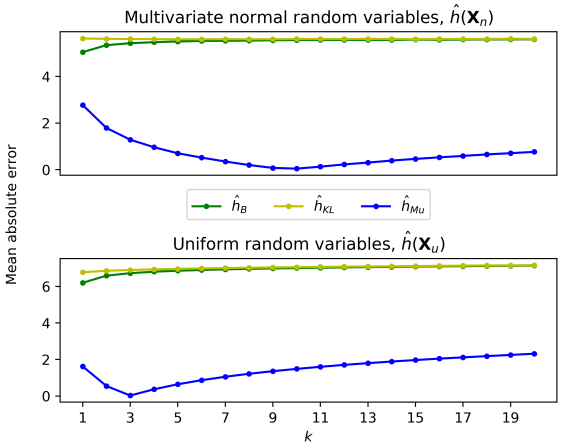


Figure 4.2: The MAE's and corresponding confidence intervals of $N = 200$ entropy estimates of random variables with multivariate normal and uniform distributions and with $k \in \{1, 2, \ldots, 20\}$. The confidence intervals are not directly visible but is e.g. $[5.60, 5.61]$ for the estimator $\hat{h}_{\mathrm{KL}}$ with multivariate normal random variables and with $k = 15$.

Figure 4.3 shows the MAE's and corresponding confidence intervals of the estimates $\hat{h}(\mathbf{X}_\mathrm{n})$ and $\hat{h}(\mathbf{X}_\mathrm{u})$ with values of $n \in \{25, 50, 75, 100, 1100, \ldots, 4{,}100, 5{,}100\}$.

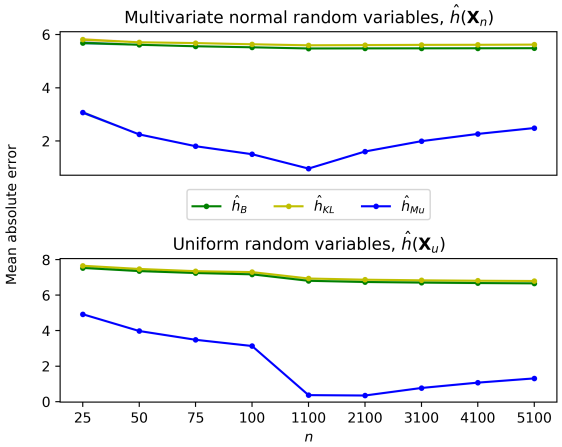


Figure 4.3: The MAE's and corresponding confidence intervals of $N = 200$ entropy estimates of random variables with multivariate normal and uniform distributions and with $n \in \{25, 50, 75, 100, 1100, \ldots, 4{,}100, 5{,}100\}$. The confidence intervals are not directly visible but is e.g. $[7.66\mathrm{e}{-01}, 7.71\mathrm{e}{-01}]$ for the estimator $\hat{h}_\mathrm{Mu}$ with uniform random variables and with $n = 3{,}100$.

As shown in Figure 4.3, the MAE is declining slowly for the estimators $\hat{h}_\mathrm{B}$ and $\hat{h}_\mathrm{KL}$. On the other hand, the estimator $\hat{h}_\mathrm{Mu}$ attains minimums at $n = 1{,}100$ (MAE of $9.65\mathrm{e}{-01}$) and at $n = 2{,}100$ (MAE of $3.44\mathrm{e}{-01}$) for the random variables with multivariate normal and uniform distributions, respectively, but the reason for the minimums being at these values is unknown.

The efficiency of the estimators is measured as the norm of the MAE's. These norms are shown in Figur 4.4.
Figure 4.4 generally show that the performances of the estimators $\hat{h}_\mathrm{B}$ and $\hat{h}_\mathrm{KL}$ are similar, whereas the performance of the estimator $\hat{h}_\mathrm{Mu}$ is always the best. When the value of k is changed, the norm of the errors of the estimator $\hat{h}_\mathrm{Mu}$ is to up to 6 times lower than those of the other estimators.

Finally, it has also been examined whether the errors are below or above the analytical values in general in order to possibly combine several estimators into a better estimator. This also provides some insights into the errors, which are not shown in the figures in this section.

For the values of d, the errors are generally above the analytical values. For the values of n and k, the errors of the estimators $\hat{h}_\mathrm{B}$ and $\hat{h}_\mathrm{KL}$ are generally above the analytical values, while the errors of the estimator $\hat{h}_\mathrm{Mu}$ is either below or above the

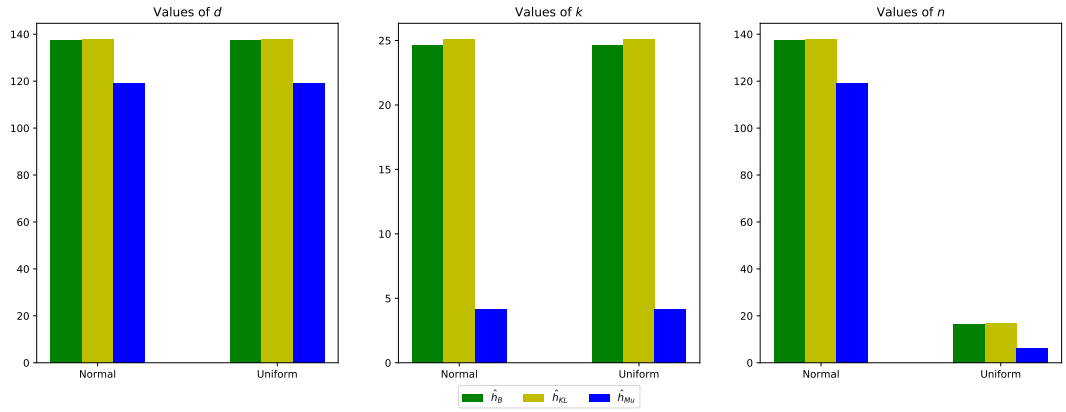


Figure 4.4: Norms of the errors for values of d (left), values of k (center), and values of n (right), where the norms of the errors for the multivariate normal and uniform random variables are to the left and right, respectively, in each plot.

analytical values at first, then becomes closer to 0, and is then either above or below for the remaining values.

From this analysis of the errors it is deemed not possible to combine several of the estimators into a better estimator.

4.1.2 Analyses of computation times

Another aspect of the estimators and how they work in an algorithm that computes the CCDI from (2.6) is how fast the entropy estimates are computed, which primarily depends on the values of d and n. It should be noted that the tests described in this section have been performed on the same computer and with no other processes running simultaneously. For reference, the computer used in the tests is a Lenovo ThinkPad T440s with an Intel® Core™ i5-4200 CPU with 1.60 GHz.

Figure 4.5 shows the mean computation times (in seconds) of the estimates $\hat{h}(\mathbf{X}_{\mathrm{n}})$ and $\hat{h}(\mathbf{X}_{\mathrm{u}})$ for the estimators with values of $d \in \{1, 5, 9, \ldots, 41, 45\}$ (left) and $n \in \{100, 1{,}100, \ldots, 4{,}100, 5{,}100\}$ (right).
According to Figure 4.5, the computation times generally seem to increase nonlinearly and are generally similar for different estimators and distributions of the random variables. However, the computations times for different values of d are sometimes different, which may be due to background processes.

The computation times of $\hat{h}_{\mathrm{B}}(\mathbf{X}_{\mathrm{n}})$ for different values of d have been fitted to the function

$$g(d) = \sqrt{a \cdot d} + b$$

where $a, b \in \mathbb{R}$ are parameters. The result is shown in Figure 4.6.

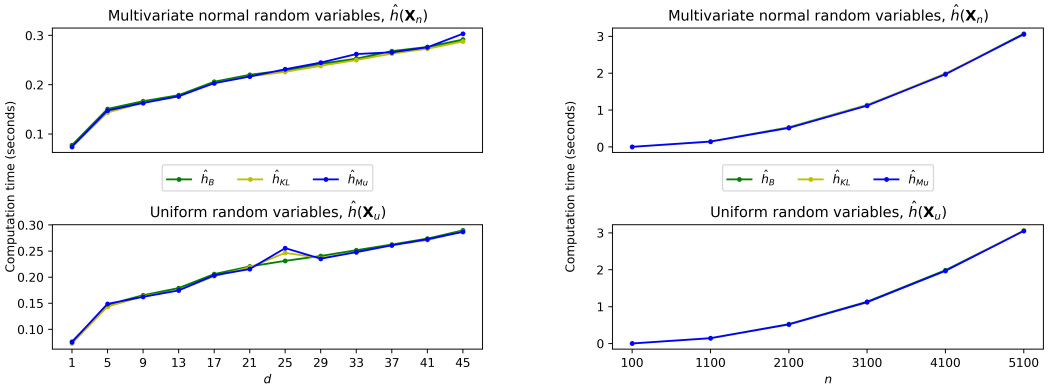


Figure 4.5: The mean computation times in seconds of $N = 200$ entropy estimates of random variables with multivariate normal and uniform distributions with $d \in \{1, 5, 9, \ldots, 41, 45\}$ (left) and $n \in \{100, 1{,}100, \ldots, 4{,}100, 5{,}100\}$ (right).

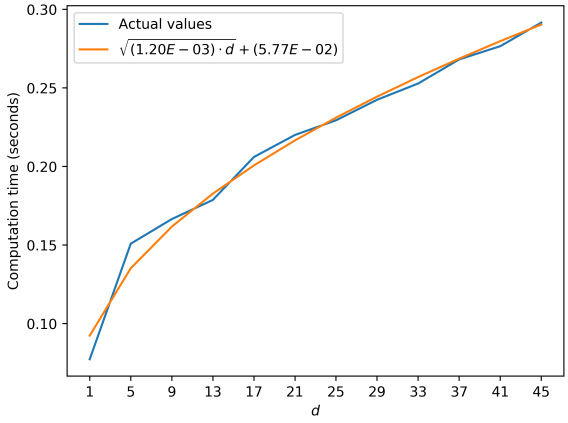


Figure 4.6: The result of fitting the function $g(d)$ to the computation times of $\hat{h}_{\mathrm{B}}(\mathbf{X}_{\mathrm{n}})$.

Figure 4.6 shows that the function $g(d)$ is a good fit to the computation times when d is changed.

The algorithmic complexities of the estimators are also analyzed in order to better understand the influence of n on the computation times. The algorithmic complexity is expressed with the big O notation in which the number of floating point operations is expressed as a function $\mathcal{O}(n)$ [10, p. 124]. As an example, the function may be expressed as $\mathcal{O}(n + n^2 + 50 \cdot n^3) = \mathcal{O}(n^3)$ where the equality means that $\mathcal{O}$ only depends on n^3 for $n \to \infty$ because it has the largest growth rate of the three terms.

This type of analysis is useful for comparing the complexities' of different algorithms to each other (after which the fastest one can be chosen) and also for identifying the time-consuming parts of the algorithms (and possibly optimizing these parts).

The first part of all of the estimators is to compute the Euclidean norms r_i between each sample $\mathbf{x}_i$, $i = 1, \ldots, n$ and all the other samples, and these distances are then sorted. The Euclidean norm $\|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}$ between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ consists of d subtractions, d squares, $d - 1$ additions, and one square root, which means that each Euclidean norm is linear in d but the Euclidean norms are computed n times. The distances are then sorted, and the complexity of this is between $\mathcal{O}(n \log_2(n))$ and $\mathcal{O}(n^2)$ in the best- and worst-case scenarios, respectively[1]. The complexity of this part is then $\mathcal{O}(dn + n^2) = \mathcal{O}(n^2)$.

The second part is different for each estimator but is mainly composed of products, divisions, sums, logarithms, and the digamma function, and the complexities of these are all of order s^2 or less, where s is the number of digits[2], which means that they are irrelevant compared to the complexity of the first part of the estimators. Collectively, the complexities of the estimators depend on the sorting algorithm and is between $\mathcal{O}(n \log_2(n))$ and $\mathcal{O}(n^2)$ in the best-case and worst-case scenarios, respectively.

Whether the computation times when n is changed behave as $\mathcal{O}(n \log_2(n))$ or $\mathcal{O}(n^2)$ is examined by fitting the computation times to the functions

$$g_1(n) = a_1 \cdot n^2 + b_1 \cdot n + c_1, \qquad g_2(n) = a_2 \cdot n \cdot \log_2(b_2 \cdot n),$$

where $a_1, b_1, c_1, a_2, b_2 \in \mathbb{R}$ are parameters. The result of fitting these functions to the computation times of $\hat{h}_{\mathrm{KL}}(\mathbf{X}_{\mathrm{n}})$ is shown in Figure 4.7 along with the values of the parameters.
Figure 4.7 shows that the second degree polynomial $g_1(n)$ fits best with the actual values.

[1]The used sorting algorithm is 'quicksort', which on average is the fastest sorting algorithm but the complexity varies. An analysis of the best- and worst-case scenarios is provided in [24].

[2]These complexities are described in [22] but this source is however questionable.

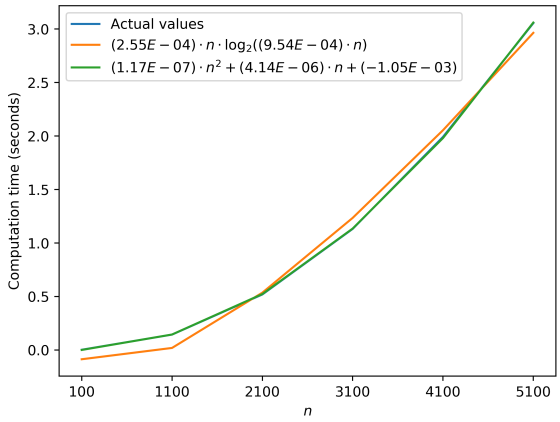


Figure 4.7: The result of fitting the functions $g_1(n)$ and $g_2(n)$ to the computation times of $\hat{h}_{\mathrm{KL}}(\mathbf{X}_{\mathrm{n}})$.

4.2 Autoregressive data

In this section, the causal conditional directed information (CCDI) from (2.6) is computed for four-dimensional synthetic data, which are dependent in both time and space. Each dimension in these data is a random process $X_i^t = \{X_i(1), \ldots, X_i(t)\}$, $i = 1, 2, 3, 4$, where the subscript denotes the number of the random process and the random variables in X_i^t are indexed by the time $t \in \mathbb{N} \setminus \{1, 2\}$ (this is a litte different from Definition 2.1.1, where the random variables in a random process are indexed by the subscript). Each of the random processes may depend on its own past (correlation in time) and/or the past of the other processes (correlation in space). Furthermore, each process contains mutually independent Gaussian white noise, $W_i(1), \ldots, W_i(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ (see Appendix A.6). The random processes used in this section[3] are

$$
\begin{aligned}
X_1(t) &= a_1 X_1(t-1) - a_2 X_1(t-2) + W_1(t), \\
X_2(t) &= b_1 X_1(t-1) + b_2 X_3(t-1) + W_2(t), \\
X_3(t) &= c_1 X_1^2(t-1) + c_2 X_2(t-1) + c_3 X_3(t-1) + W_3(t), \\
X_4(t) &= d_1 X_1(t-1) + d_2 X_3(t-1) + d_3 X_4(t-1) + W_4(t), \quad t \in \mathbb{N} \setminus \{1, 2\},
\end{aligned} \tag{4.1}
$$

where the degree of dependence in time and/or space is controlled by the vectors of parameters $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$.

Note that $X_1(t)$ is only dependent in time but the other processes are all dependent on it in space. Note also that it is important to choose the parameters $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ properly in order to ensure that the random processes are stationary (see Appendix A.6). Finally, the first fifth part of each process is considered a transcient period and

[3] These have been suggested by Payam Baboukani from AAU but have however been modified a little bit by the author.

is removed in order for the process to be steady state (which means that the behavior of the process does not change).

Three cases of these autoregressive data are considered with different parameters, where the parameters not described are just 0, which e.g. in the first case means that $X_3(t)$ and $X_4(t)$ are just mutually independent Gaussian white noise processes.

1. In the first case, only $X_1(t)$ and $X_2(t)$ from (4.1) are dependent with $a_1 = 0.8$.
2. The second case is similar to the first case but now $X_3(t)$ is dependent on $X_2(t)$ through $b_2 = 0.7$, $c_1 = 0.6$, and $c_3 = 0.8$.
3. The third case is similar to the second case but now $X_4(t)$ is also dependent on $X_3(t)$ through $d_2 = 0.9$.

The data are used to calculate the CCDI (see (2.6)), where $\mathbf{X}^n$ is $X_1(t)$, $\mathbf{Y}^n$ is $X_2(t)$, and $\mathbf{Z}^n$ is a vector of $X_3(t)$ and $X_4(t)$ in all three cases. Furthermore, the spatial correlation between $X_1(t)$ and the remaining processes with parameters different from 0 is in all three cases controlled by b_1, and the CCDI is therefore computed for every value $b_1 \in \mathbf{b}_1 = \{-1.0, -0.9, \ldots, 0.9, 1.0\}$.

The main idea with this test is to check whether the system is able to compute the CCDI reasonably. More complex cases (with more parameters different from 0) could of course be considered but it is necessary to be able to compare the computed CCDI with how the processes are dependent and how the CCDI is expected to behave when the value of b_1 is changed. The expected result is generally in all three cases that the CCDI will be low for b_1 close to 0 since there will be no flow of information between $X_1(t)$ and the remaining processes and that the CCDI will grow for b_1 going towards ± 1.

In each of the three cases with the described parameter values and with b_1 equal to each of the values in $\mathbf{b}_1$, the CCDI has been computed $N = 100$ times after which the mean of the computed CCDI's is found. In order to ensure that the processes used in the test are stationary, each of the processes are tested for stationarity with the ADF test (see Appendix A.6) before being used, and if either of the processes are not stationary, new processes are generated and tested for stationarity before being used[4]. The results are shown with the confidence intervals (see Appendix A.5) for the estimator $\hat{h}_{\mathrm{Mu}}$ in Figure 4.8.

[4]In total, 1 process was nonstationary in case 2, and 6 processes were nonstationary in case 3.

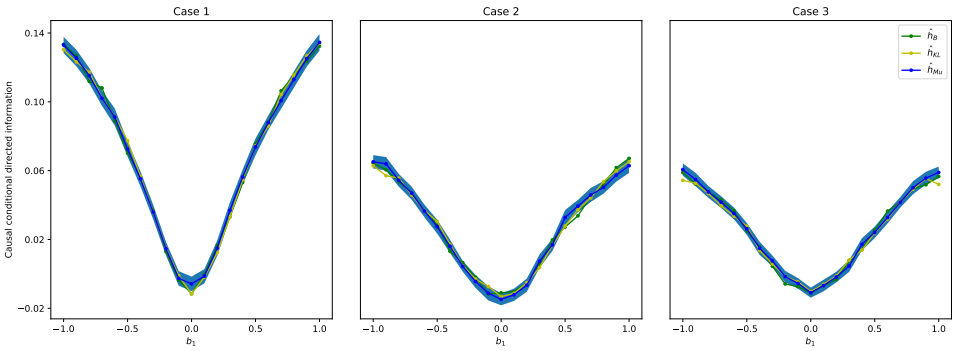


Figure 4.8: The CCDI in the three cases with $b_1 \in \mathbf{b}_1$ and with the confidence intervals for the estimator $\hat{h}_{\mathrm{Mu}}$.

The results in Figure 4.8 is generally as expected since the CCDI is highest for $b_1 = \pm 1$ and then gradually decreases to its minimum close to 0 at $b_1 = 0$. The CCDI is lower in case 2 and 3, which is expected since the CCDI will be lower if the correlation with the process that is causally given is higher. In case 1, $\mathbf{Z}^n$ is Gaussian white noise, whereas the random variables are spatially correlated with $X_1(t)$ and $X_2(t)$ through b_2 and c_1 in case 2 and 3. The estimates computed by the estimators are generally similar but however with minor deviations.

It is asserted that the CCDI is computed correctly and that the estimator being used only has a minor influence on these computations.

4.3 Analyses of EEG data

In this section, the CCDI (see (2.6)) is computed for actual EEG data, which are obtained online [2].

The data contains measurements of 20 subjects, and the measurements of each subject consists of 5 sessions with closed eyes and 5 sessions with open eyes (i.e. 200 sessions in total) with each session lasting for 10 seconds. The data are measured with a cap equipped with 16 electrodes with locations FP1, FP2, FC5, FC6, FZ, T7, CZ, T8, P7, P3, PZ, P4, P8, O1, Oz, and O2 according to the 10-10 international system [16, p. 419], where each of these electrodes corresponds to a dimension.

The sampling frequency is 512 samples per second, which means that each session should contain 5,120 samples. But the mean number of samples in each session is 8,180.19, and 11 sessions contain more than 15,000 samples. It can not be assumed that these sessions with an excess of samples only contain EEG measurements where the eyes e.g. have been closed all the time so these measurements should be managed in some way. In the description of the experiment in [2, p. 5], it says that the subject was asked to either open or close their eyes before each session, and the row with the timestamp corresponding to the beginning of the session is marked in the

dataset. Therefore, each session can simply be truncated to only contain the 5,120 samples from the beginning of the session since these samples must correspond to the 10 seconds of the actual session (the remaining samples can possibly be considered as breaks between the sessions). A histogram with the number of samples in all of the sessions (before being truncated to only contain 5,120 samples) is shown in Figure B.1 in Appendix B. Note that three sessions are shorter than 5,120 samples (with 4,822, 5,013, and 5,116 samples) but these sessions are not altered in any way.

As described in Chapter 1, research suggests that there is a big difference in the connectivity between the occipital and frontal areas of the brain when the eyes are closed and opened. The objective in this section is to compute the CCDI between the occipital and frontal areas when the eyes are closed and opened and assert the difference between these computations.

Figure 4.9 shows colormaps of the CCDI between the occipital and frontal areas when the eyes are closed (left) and opened (right). However, only the first 50 out of the total 100 sessions with either closed or opened eyes have been used in these analyses because it is quite time-consuming to compute the CCDI.

Note that in Figure 4.9, $\mathbf{X}^n$ and $\mathbf{Y}^n$ from (2.6) is only one electrode each (on the first and second axis, respectively), and $\mathbf{Z}^n$ is the remaining six electrodes. Furthermore, the CCDI between the same electrodes are not computed and just manually set to 0. Finally, the results in Figure 4.9 has only been computed with the estimator $\hat{h}_{\mathrm{KL}}$ since only minor deviations between the computations of the CCDI in Section 4.2 were seen.

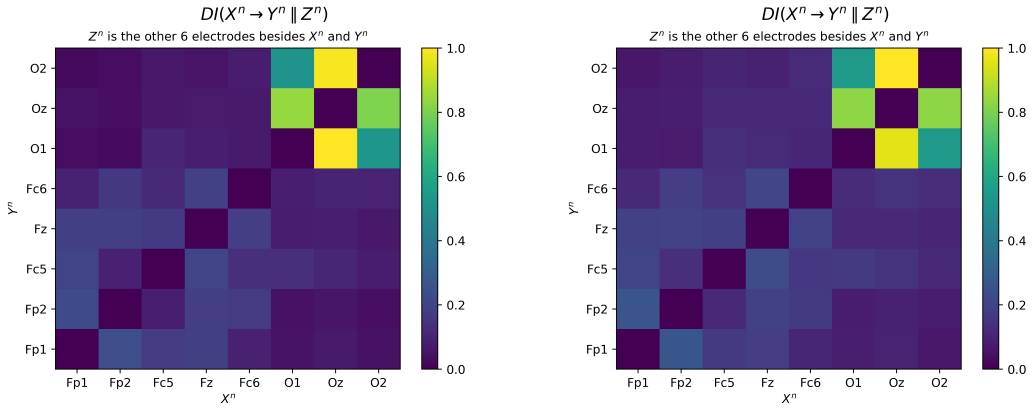


Figure 4.9: The CCDI between each of the eight electrodes belonging to the frontal and occipital areas. The included data are 50 of the sessions closed eyes (left) and opened eyes (right).

No immediate differences between the CCDI's with opened and closed eyes can be seen in Figure 4.9. The absolute differences between the CCDI's are shown in Figure 4.10.

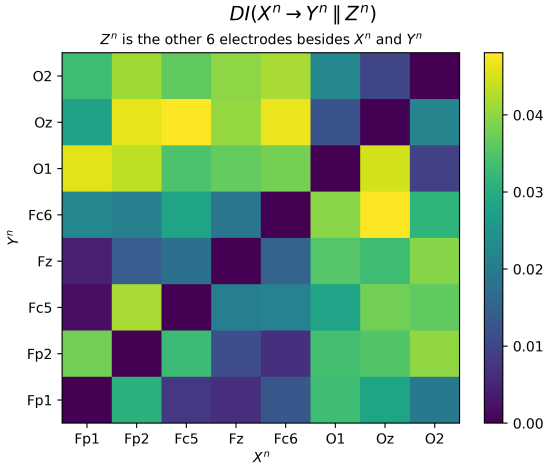


Figure 4.10: The absolute differences between the CCDI with opened and closed eyes.

Figure 4.10 shows that there only are minor differences in the CCDI when the eyes are closed compared to when they are opened.

5 | Discussion

In this chapter, the research question from Section 1.2 is discussed based on both the theoretical and experimental results. The research question and associated study questions from Section 1.2 are:

How can kNN-based entropy estimators be used to estimate the causal conditional directed information measure, which quantifies the information flow between different parts of the brain?

1. What is the underlying theory of the directed information measure?
2. How can the CCDI be computed sequentially without using all of the available data?
3. How can well-known kNN-based entropy estimators be derived?
4. Is it possible to derive an improved estimator?
5. How can the implemented estimators be tested on both synthetic and actual data?

The methods used to answer these questions are firstly discussed in Section 5.1, the results are then discussed in Section 5.2, and the chapter is concluded by a general assessment based on the research question.

5.1 Methods

The directed information measure that has been used in this thesis is the causal conditional directed information (CCDI), and the underlying theory is described in Chapter 2. However, different directed information measures such as the transfer entropy, the directed information (which is also defined in Definition 2.2.3), and the momentary information transfer are described in [21, pp. 7,28,30] and could have been used instead of the CCDI. The difference between these directed information measures and the CCDI is the *causal conditioning* in the CCDI, which means that the past of both random processes are considered. Therefore, the CCDI is a favourable measure of directed information when considering the information flow in the brain, which simultaneously is affected by different parts and where it is necessary to determine if there is a causal relationship.

Since all of the random variables are not necessarily dependent on each other, it is not necessary to use them all in the definition of the CCDI in (2.5). Instead, the

random variable $\mathbf{X}_i$ has been assumed to only be dependent on the previous l random variables, which means that the number of terms in the sum can be greatly reduced (depending on the number of random variables), which means that the computation time of the CCDI can also be reduced (see (2.6)).

Several well-known kNN-based entropy estimators have been derived in this thesis by among others using rather simple probability theory and results on integration. These estimators are shortly described in the following.

1. The estimator $\hat{h}_{\mathrm{B}}$ is a simple estimator and is derived from the assumption that the density inside the ball $\mathcal{B}(\mathbf{x}_1, r_1)$ is constant.

2. The estimator $\hat{h}_{\mathrm{S}}$ is derived by showing that the estimator $\hat{h}_{\mathrm{B}}$ is asymptotically biased and by then subtracting its asymptotic mean from it, which makes $\hat{h}_{\mathrm{S}}$ asymptotically unbiased. However, it turns out that the term making $\hat{h}_{\mathrm{B}}$ asymptotically biased is below 1 for $k = 1$ and 0 for larger values of k.

3. The estimator $\hat{h}_{\mathrm{KL}}$ is derived by using a key result (see (3.15)), which it has not been possible to prove is true, and by also assuming that the density inside the ball $\mathcal{B}(\mathbf{x}_1, r_1)$ is constant. This estimator is furthermore seen to be asymptotically equal to the estimator $\hat{h}_{\mathrm{S}}$, which means that it is also asymptotically unbiased, and which means that it is also equal to $\hat{h}_{\mathrm{B}}$ for larger values of k.

4. The estimator $\hat{h}_{\mathrm{Mu}}$ is derived by performing numerical integration on the integral in the aforementioned key result in (3.15) since it did not seem possible to derive the result analytically. The numerical integration is performed on a grid of different values of n, d but the resulting values of the integral is observed not to vary considerably for these values of n, d, and a constant is used instead.

5.2 Results

The implemented estimators are tested on both synthetic and actual data in Chapter 4.

The estimators are firstly used to compute entropy estimates of random variables with multivariate normal and uniform distributions for which the analytical values of the entropies have been derived in Chapter 2. The mean absolute errors of the estimates are analyzed for different values of d, k, and n in Section 4.1.1.

1. The analysis of different values of d shows that the relationship between d and the error is almost perfectly linear, which means that the estimators only compute the same errors in each dimension.

2. The analysis of different values of k shows that the errors are almost constant for the estimators $\hat{h}_{\mathrm{B}}$ and $\hat{h}_{\mathrm{KL}}$, whereas the estimator $\hat{h}_{\mathrm{Mu}}$ performs much better. The estimator $\hat{h}_{\mathrm{Mu}}$ attains minimums at $k = 10$ for the multivariate normal random variables (MAE of 4.52e−02) and at $k = 3$ for the uniform random variables (MAE of 2.62e−02).

3. For the analysis of different values of n, the errors are also almost constant for the estimators $\hat{h}_{\mathrm{B}}$ and $\hat{h}_{\mathrm{KL}}$, whereas the estimator $\hat{h}_{\mathrm{Mu}}$ performs much better. The estimator $\hat{h}_{\mathrm{Mu}}$ attains minimums at $n = 1{,}100$ for the multivariate normal random variables (MAE of 9.65e−01) and at $n = 2{,}100$ for the uniform random variables (MAE of 3.44e−01).

Further analyses of the errors also shows that the errors of the estimator $\hat{h}_{\mathrm{Mu}}$ is either above or below the analytical values at first, then attains its minimum close to an MAE of 0 after which the error is below or above the analytical values. On the other hand, the errors of the estimators $\hat{h}_{\mathrm{B}}$ and $\hat{h}_{\mathrm{KL}}$ are generally always above the analytical values (except for low values of d).
The efficiency of the estimators is also analyzed as the norms of the errors. The norms of the errors for the estimator $\hat{h}_{\mathrm{Mu}}$ are always below those of the estimators $\hat{h}_{\mathrm{B}}$ and $\hat{h}_{\mathrm{KL}}$, and they are up to 6 times lower than these norms in the analyses of different values of k. This means that $\hat{h}_{\mathrm{Mu}}$ is the best estimator, and its derivation is claimed to be a key result in this thesis.

Other options for the distributions of the random variables used in these tests could also have been considered, but the multivariate normal and uniform distributions have been used because the expressions for the analytical values are fairly simple and because all of the estimators assume that the density is constant. Unfortunately, it has not been possible to implement an estimator that assumes a normal density (see Chapter 7) in this thesis due to time limitations.

If one has the option to choose the values of n and d, one should also consider the computation times.

An analysis of the algorithmic complexities of the implemented estimators shows that the complexity is between order $\mathcal{O}(n \log_2(n))$ and $\mathcal{O}(n^2)$ in the best- and worst-case scenarios. The computation times for different values of n are also seen to follow a second-degree polynomial (see Figure 4.7). Furthermore, the computation times for different values of d are seen to be of order $\mathcal{O}(\sqrt{d})$ (see Figure 4.6).

In Section 4.2, the estimators were used to compute the CCDI between autoregressive data with four specified random processes with both autocorrelation (correlation within the process) and spatial correlation (correlation between the processes). In these tests, the parameters of the random processes are chosen such that a single parameter controls the correlation between the first process and the other processes,

which means that expectations for how the CCDI to behave can be made. The CCDI is seen to follow these expectations in all three cases. Furthermore, the computations of the CCDI were almost identical for the estimators, which means that the errors of the estimators actually only have a little significance when computing the CCDI.

In Section 4.3, the estimator $\hat{h}_{\mathrm{KL}}$ was used to compute the CCDI of actual EEG data. Note that only the estimator $\hat{h}_{\mathrm{KL}}$ was used to compute the CCDI due to the aforementioned minor differences between the computations of the CCDI of the autoregressive data in Section 4.2. Furthermore, the estimator $\hat{h}_{\mathrm{Mu}}$ was not used even though its errors were seen to be smallest in Section 4.1.1 because it was only developed shortly before the end of the project period.

The used EEG data consists of sessions with both opened and closed eyes, and the CCDI between the occipitial and frontal areas of the brain has been computed when the eyes are opened and closed in order to assess the difference between these because research suggests that the connectivity is higher when the eyes are closed. However, only 50 of the 100 sessions with either closed or opened eyes were used in these analyses because it is very time-consuming to compute the CCDI.

The absolute differences between the CCDI's with the eyes opened and closed are shown in Section 4.10, and there are only minor differences between the CCDI's. Therefore, further analyses are required to determine if there is a difference between the occipital and frontal areas of the brain when the eyes are closed compared to when they are opened. Such analyses could e.g. include only analyzing a subset of the electrodes that have been used in the analyses described in this thesis.

It is generally assessed that kNN-based entropy estimators have been used to estimate the causal conditional directed information measure in this thesis. This involves describing the theory of this measure and how to compute it sequentially as well as deriving different kNN-based entropy estimators, which have been tested on synthetic and actual data. One of the derived estimators is constructed by the author, and the norm of its errors is seen to be up to 6 times lower than those of the other estimators. However, which estimator is being used when computing the CCDI only has a minor influence on the resulting values of the CCDI. Finally, only minor differences between the CCDI's with opened and closed eyes are seen, contrary to what is seen in the literature.

6 | Conclusion

This thesis considers how to reliably estimate the causal conditional directed information with kNN-based entropy estimators.

It is firstly shown that the causal conditional directed information can be written as a sum of joint entropies, which means that it can be estimated with entropy estimators. It is then assumed that each random variable only depend on a number of the previous random variables, which makes it possible to reduce the number of terms and therefore also reduce the computation time.

The first estimator derived in this thesis is a basic estimator denoted by $\hat{h}_{\mathrm{B}}$, which is also described in the literature. Another estimator, which is denoted by $\hat{h}_{\mathrm{KL}}$, is shown to be an asymptotically unbiased version of $\hat{h}_{\mathrm{B}}$ but they are however equal for large values of k and n. It was found that a key result was missing from the derivation of the estimator $\hat{h}_{\mathrm{KL}}$ in the literature. This result was instead estimated with numerical integration, which resulted in a new estimator denoted by $\hat{h}_{\mathrm{Mu}}$.

The estimators have been tested on synthetic data in order to assess their performances. In these tests, the dimensions d, the numbers of neighbors k, and the numbers of samples n are changed one at a time. It is generally seen that the performances of the estimators $\hat{h}_{\mathrm{B}}$ and $\hat{h}_{\mathrm{KL}}$ are similar, whereas the new estimator $\hat{h}_{\mathrm{Mu}}$ is up to 6 times better when the norms of the estimators' errors are compared, which makes this estimator a key result in this thesis.

The estimators have then been used to compute the causal conditional directed information of autoregressive data, where the expectations to the behaviour of the causal conditional directed information can be made. The computations of the causal conditional directed information are generally seen to follow these expectations, and there is furthermore only minor differences between the computations from the three estimators – in spite of the large differences between the errors of the estimators.

One of the estimators is finally used to compute the causal conditional directed information between different electrodes from actual EEG data. The hypothesis has been that the occipitial and frontal areas of the brain have a stronger connectivity when the eyes are closed compared to when they are opened, which is examined by computing the causal conditional directed information between these areas when the eyes are closed and opened and then examining the differences.

The results show that there only is a minor difference in the CCDI's when the eyes are closed and opened. Further studies with only a subset of the electrodes used in this thesis are deemed necessary in order to assess whether there is a difference.

7 | Future studies

This chapter considers another estimator, which it has not been possible to implement in this thesis due to time limitations.

According to [11, p. 2], the errors for the estimator $\hat{h}_{\text{KL}}$ are large when the dimensionality is high and when the data are highly correlated, which primarily is due to the assumption of a constant probability density in $\mathcal{B}(\mathbf{x}_1, r_1)$. The solution suggested by [11] is to represent the density in $\mathcal{B}(\mathbf{x}_1, r_1)$ as proportional to a Gaussian function.

The proposed estimator described in [11, pp. 2-3] assumes that the pdf of the p nearest neighbors of $\mathbf{x}_1$ for $p \geq k$, which are denoted by $\tilde{\mathbf{x}} \in \mathbb{R}^{p \times d}$, is proportional to a Gaussian function

$$f_G(\mathbf{x}) \approx \rho \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{\tilde{\mathbf{x}}}\right)^{\mathsf{T}} \mathbf{S}_{\tilde{\mathbf{x}}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\tilde{\mathbf{x}}}\right)\right), \tag{7.1}$$

where $\boldsymbol{\mu}_{\tilde{\mathbf{x}}} \in \mathbb{R}^p$ and $\mathbf{S}_{\tilde{\mathbf{x}}} \in \mathbb{R}^{p \times p}$ are the mean vector and covariance matrix of $\tilde{\mathbf{x}}$, respectively, and where ρ is the proportionality constant. The expression for the pdf in (7.1) can also be written as

$$f_G(\mathbf{x}) \approx f_G(\mathbf{x}_1)\frac{g(\mathbf{x})}{g(\mathbf{x}_1)}, \quad g(\mathbf{x}) = \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{\tilde{\mathbf{x}}}\right)^{\mathsf{T}} \mathbf{S}_{\tilde{\mathbf{x}}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\tilde{\mathbf{x}}}\right)\right).$$

By Definition 3.1.2, $F(r_1|\mathbf{x}_1)$ can then be written as:

$$\begin{aligned} F(r_1|\mathbf{x}_1) = \int_{\|\mathbf{x}-\mathbf{x}_1\|<r_1} f_G(\mathbf{x})\,\mathrm{d}\mathbf{x} &= f_G(\mathbf{x}_1)\frac{1}{g(\mathbf{x}_1)}G(r_1|\mathbf{x}_1), \\ G(r_1|\mathbf{x}_1) &= \int_{\|\mathbf{x}-\mathbf{x}_1\|<r_1} g(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{7.2}$$

By comparing this result to the expression in (3.16), it is seen that $\eta_1 = \frac{1}{g(\mathbf{x}_1)}G(r_1|\mathbf{x}_1)$, which is combined with the result in (3.17) to obtain the entropy estimator $\hat{h}_{\text{kpn}}(\mathbf{X})$ (where the result in (3.15) is assumed to be true) [11, p. 3]

$$\begin{aligned} \hat{h}_{\text{kpn}}(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{1}{g(\mathbf{x}_i)}G(r_i|\mathbf{x}_i)\right) + \psi(n) - \psi(k) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\ln(G(r_i|\mathbf{x}_i)) - \ln(g(\mathbf{x}_i))\right) + \psi(n) - \psi(k), \end{aligned} \tag{7.3}$$

where 'kpn' refers to the parameters k, p, and n. It is however difficult to evaluate $G(r_i|\mathbf{x}_i)$, especially in high dimensions. In [11, p. 3] the numerical method *expectation propagation for multivariate Gaussian probabilities* (EPMGP) is suggested, but it has

however not been possible to implement this method due to time limitations.

The estimator $\hat{h}_{\text{kpn}}$ is similar to $\hat{h}_{\text{KL}}$ since it also relies on the key result in (3.15). Therefore, another estimator, where the constant M_{n} from the numerical integration in Section 3.3.1 is used instead of the result in (3.15) (similarly to the difference between the estimators $\hat{h}_{\text{KL}}$ and $\hat{h}_{\text{Mu}}$ as described in Section 3.4), can then be written as

$$\hat{h}_{\text{Mn}} = \frac{1}{n}\sum_{i=1}^{n}\left(\ln(G(r_i|\mathbf{x}_i)) - \ln(g(\mathbf{x}_i))\right) + M_{\text{n}}.$$

Bibliography

[1] Amblard, Pierre-Olivier and Michel, Olivier J. J. "On directed information theory and Granger causality graphs". In: *Journal of Computational Neuroscience* 30.1 (Mar. 2010), pp. 7–16. ISSN: 1573-6873. DOI: 10.1007/s10827-010-0231-x.

[2] CATTAN, Grégoire, Coelho Rodrigues, Pedro Luiz, and Congedo, Marco. *EEG Alpha Waves Dataset*. Research Report. GIPSA-LAB, Dec. 2018. URL: https://hal.archives-ouvertes.fr/hal-02086581.

[3] Chen, Yen-Chi. *Lecture 7: Density Estimation: k-Nearest Neighbor and Basis Approach*. Visited 06-01-2020. 2018. URL: http://faculty.washington.edu/yenchic/18W_425/Lec7_knn_basis.pdf.

[4] Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory*. 2nd ed. ISBN-13: 978-0-471-24195-9. John Wiley & Sons, Inc., 2006.

[5] Dekking, F. M. et al. *A Modern Introduction to Probability and Statistics*. Springer, 2005. ISBN: 978-1-85233-896-1.

[6] Edwards, C. Henry and Penney, David E. *Calculus – Early Transcendentals*. 7th ed. Pearson, 2014. ISBN: 978-0-321-99838-5.

[7] Geller, Aaron S. et al. "Eye closure causes widespread low-frequency power increase and focal gamma attenuation in the human electrocorticogram". In: *Clinical Neurophysiology* 125.9 (Sept. 2014), pp. 1764–1773. DOI: 10.1016/j.clinph.2014.01.021.

[8] Kay, Steven. *Intuitive probability and random processes using MATLAB ®*. Jan. 2012. ISBN: 978-0-387-24157-9. DOI: 10.1007/b104645.

[9] Kraskov, Alexander, Stögbauer, Harald, and Grassberger, Peter. "Estimating mutual information". In: *Physical Review E* 69 (6 June 2004), p. 066138. DOI: 10.1103/PhysRevE.69.066138. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

[10] Larsen, Torben, Arildsen, Thomas, and Jensen, Tobias L. *Behavioral Simulation and Computing for Signal Processing Systems*. Wiley-Blackwell, 2010.

[11] Lombardi, Damiano and Pant, Sanjay. "Nonparametric k-nearest-neighbor entropy estimator". In: *Phys. Rev. E* 93.1 (Jan. 2016), p. 013310. DOI: 10.1103/PhysRevE.93.013310. URL: https://link.aps.org/doi/10.1103/PhysRevE.93.013310.

[12] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.25 of 2019-12-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. URL: http://dlmf.nist.gov/.

[13] *NIST/SEMATECH e-Handbook of Statistical Methods – Shapiro-Wilk test*. Visited 03-06-2020. URL: https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm.

[14] Olofsson, Peter and Andersson, Mikael. *Probability, Statistics, and Stochastic Processes*. 2nd ed. Wiley, 2012.

[15] Pishro-Nik, Hossein. *Normal (Gaussian) Distribution*. Visited 25-05-2020. URL: https://www.probabilitycourse.com/chapter4/4_2_3_normal.php.

[16] Schoenberg, Mike R., Werz, Mary A., and Drane, Daniel L. "Epilepsy and Seizures". In: *The Little Black Book of Neuropsychology*. Ed. by Schoenberg, Mike R. and Scott, James G. Springer, 2011. Chap. 16, pp. 423–519. ISBN: 9781118456989. DOI: 10.1007/978-0-387-76978-3.

[17] Shumway, Robert H. and Stoffer, David S. *Time Series Analysis and Its Applications*. Ed. by DeVeaux, Richard, Fienberg, Steven, and Olkin, Ingram. 4th ed. Springer, 2017. DOI: 10.1007/978-3-319-52452-8.

[18] Singh, Harshinder et al. "Nearest Neighbor Estimates of Entropy". In: *American Journal of Mathematical and Management Sciences* 23.3-4 (Aug. 2003), pp. 301–321. DOI: 10.1080/01966324.2003.10737616. URL: https://doi.org/10.1080/01966324.2003.10737616.

[19] Tan, Bo et al. "The Difference of Brain Functional Connectivity between Eyes-Closed and Eyes-Open Using Graph Theoretical Analysis". In: *Computational and Mathematical Methods in Medicine* (2013). DOI: 10.1155/2013/976365.

[20] Wei, Jie et al. "Eyes-Open and Eyes-Closed Resting States With Opposite Brain Activity in Sensorimotor and Occipital Regions: Multidimensional Evidences From Machine Learning Perspective". In: *Frontiers in Human Neuroscience* 12 (2018), p. 422. DOI: 10.3389/fnhum.2018.00422.

[21] Wibral, Michael, Lizier, Joseph T., and Vicente, Raul, eds. *Directed Information Measures in Neuroscience*. Understanding Complex Systems. Springer, 2014. DOI: 10.1007/978-3-642-54474-3.

[22] Wikipedia. *Computational complexity of mathematical operations*. Visited 17-05-2020. URL: https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations.

[23] Wikipedia. *Digamma function, Inequalities*. Visited 21-05-2020. URL: https://en.wikipedia.org/wiki/Digamma_function#Inequalities.

[24] Wikipedia. *Quicksort*. Visited 16-05-2020. URL: https://en.wikipedia.org/wiki/Quicksort.

[25] Xiong, Wanting, Faes, Luca, and Ivanov, Plamen Ch. "Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations". In: *Phys. Rev. E* 95 (June 2017), p. 062114. DOI: 10.1103/PhysRevE.95.062114. URL: https://link.aps.org/doi/10.1103/PhysRevE.95.062114.

A | Additional definitions and results

In this appendix some additional definitions and results are shown.

A.1 Definitions from probability theory

For a continuous random variable $\mathbf{X} \in \mathbb{R}^d$, the function $F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$[1] is called the cumulative distribution function (cdf) of $\mathbf{X}$, and the function $f(\mathbf{x}) = F'(\mathbf{x})$ is called the probability density function (pdf)[2] of $\mathbf{X}$. The function $f(\mathbf{x})$ is only a possible pdf of $\mathbf{X}$ if [14, p. 86]

$$f(\mathbf{x}) \geq 0 \quad \text{and} \quad \int_{\mathbb{R}^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1 \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^d. \tag{A.1}$$

The cdf can also be computed from the pdf as [14, p. 85]

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{t})\,\mathrm{d}\mathbf{t}, \quad \mathbf{t} \in \mathbb{R}^d.$$

For n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the cdf can be estimated as $\hat{F}$ through the empirical distribution function (edf) [14, p. 325]

$$\hat{F}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\mathbf{x}_i \leq \mathbf{x}), \tag{A.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function

$$\mathbb{I}(\mathbf{x} \leq \mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \leq \mathbf{y} \\ 0 & \text{otherwise.} \end{cases}$$

For a continuous random variable $\mathbf{X} \in \mathbb{R}^d$ with pdf $f(\mathbf{x})$, the expected value is [14, p. 98]

$$\mathbb{E}[\mathbf{X}] = \int_{\mathbb{R}^d} \mathbf{x} f(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{A.3}$$

[1] In general, $P(\cdot)$ denotes the probability of an event.

[2] The function $p(\mathbf{x}_k) = P(\mathbf{X} = \mathbf{x}_k)$, $k = 1, 2, \ldots$ for a discrete random variable $\mathbf{X}$ is similarly called the probability mass function (pmf).

For another continuous random variable $\mathbf{Y} \in \mathbb{R}^d$, the expected value of $\mathbf{Y}$ given that $\mathbf{X} = \mathbf{x}$ can be written as [14, p. 186]

$$\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \int_{\mathbb{R}^d} \mathbf{y} f(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y}. \tag{A.4}$$

A.2 Probability distributions

The Poisson distribution is defined in Definition A.2.1 [14, p. 117].

Definition A.2.1 *Let $\mathbf{X} \in \mathbb{R}^d$ be a discrete random variable with probability mass function*

$$p(k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

Then $\mathbf{X}$ is said to have a Poisson distribution with the parameter λ, which is written as $\mathbf{X} \sim \mathrm{Poi}(\lambda)$.

The binomial distribution is defined in Definition A.2.2 [14, p. 112].

Definition A.2.2 *Let $\mathbf{X} \in \mathbb{R}^d$ be a discrete random variable with probability mass function*

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n.$$

Then $\mathbf{X}$ is said to have a binomial distribution with the parameters n, p, which is written as $\mathbf{X} \sim \mathrm{Bin}(n, p)$. The coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is known as the binomial coefficient.

Note that for the binomial distribution, there are only two possible outcomes, which have probabilities p and $1 - p$, respectively, and the number of these outcomes are k and $n - k$, respectively. Furthermore, these outcomes are i.i.d. Finally, the expected value of the binomial distribution is $\mathbb{E}[\mathbf{X}] = np$ [14, pp. 112-113].

The normal distribution is defined in Definition A.2.3 [14, p. 127].

Definition A.2.3 *Let $\mathbf{X} \in \mathbb{R}^d$ be a random variable with pdf*

$$f(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu})^2\right),$$

where μ and σ^2 is the expected value and variance of $\mathbf{X}$, respectively. Then $\mathbf{X}$ is said to have a normal distribution with parameters μ and σ, which is written as $\mathbf{X} \sim \mathcal{N}(\mu, \sigma)$.

A random variable is further said to have a standard normal distribution if $\mathbf{X} \sim \mathcal{N}(0,1)$ [14, p. 128].

The multivariate normal distribution is defined in Definition A.2.4 [14, pp. 226-227].

Definition A.2.4 *Let* $\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_n]^\mathsf{T}$ *be a vector of random variables with* $\mathbf{X}_i \in \mathbb{R}^d$ *for* $i = 1, \ldots, n$ *and let* $\mathbf{x} = [\mathbf{x}_1 \cdots \mathbf{x}_n]^\mathsf{T}$ *be a vector of realizations of* $\mathbf{X}$*. Furthermore, let* $\boldsymbol{\mu} = [\mu_1 \cdots \mu_n]^\mathsf{T}$ *be the mean vector with* $\mu_i = \mathbb{E}[\mathbf{X}_i]$ *and let* $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ *be the covariance matrix, which is assumed to be invertible and for which*

$$\Sigma_{ij} = \mathbb{E}[(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)], \quad 1 \leq i, j \leq n.$$

If $\mathbf{X}$ *has an* n*-dimensional joint pdf*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

then $\mathbf{X}$ *is said to have a multivariate normal distribution with parameters* $\boldsymbol{\mu}, \boldsymbol{\Sigma}$*, which is written as* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*.*

Note that the $(i,j)^{\text{th}}$ entrance of the covariance matrix in Definition A.2.4, Σ_{ij}, can be written as:

$$\begin{aligned}\Sigma_{ij} &= \mathbb{E}[(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)] = \mathbb{E}[\mathbf{X}_i\mathbf{X}_j - \mathbf{X}_j\mu_i - \mu_j\mathbf{X}_i + \mu_j\mu_i] \\ &= \mathbb{E}[\mathbf{X}_i\mathbf{X}_j] - \mu_i\mathbb{E}[\mathbf{X}_j] - \mu_j\mathbb{E}[\mathbf{X}_i] + \mu_j\mu_i = \mathbb{E}[\mathbf{X}_i\mathbf{X}_j] - \mu_i\mu_j \triangleq \mathrm{C}_{\mathbf{X}}(i,j),\end{aligned}$$

which is known as the covariance between $\mathbf{X}_i$ and $\mathbf{X}_j$ [14, p. 197].

A.3 The Poisson approximation to the binomial distribution

The Poisson approximation to the binomial distribution is written in Proposition A.3.1 and proven afterwards.

Proposition A.3.1 *For* $n \to \infty$*,* $p \to 0$*, and* $np \to \lambda$*, the binomial distribution with parameters* n, p *converges to the Poisson distribution with parameter* λ*, i.e.*

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np\to\lambda}} \binom{n}{k} p^k(1-p)^{n-k} = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

Proof Since it is required that $np \to \lambda$, the parameter p is written as $\frac{\lambda}{n}$ in the probability mass function for the binomial distribution such that

$$p(k) = \binom{n}{k} p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!} \cdot p^k(1-p)^{n-k}$$

$$
\begin{aligned}
&= \frac{n \cdot (n-1) \cdot (n-2) \cdots 1}{(n-k) \cdot (n-k-1) \cdot (n-k-2) \cdots (n-k-(n-k)+1) \cdot k!} \\
&\quad \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{n} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k},
\end{aligned}
$$

where the terms in the denominator in the second line cancel out with the corresponding terms in the numerator. Taking the limit $n \to \infty$ means that the first k fractions that depend on n and the term $\left(1 - \frac{\lambda}{n}\right)^{-k}$ all converges to 1.

Consider the part $\left(1 - \frac{\lambda}{n}\right)^n$ of the expression. The function $\ln(1+z)$ can be written as

$$
\ln(1+z) = \int_0^z \frac{1}{1+t}\,\mathrm{d}t,
$$

where $\frac{1}{1+t} = 1 - t + t^2 - t^3 + \cdots$ with $t < |1|$ since

$$
\begin{aligned}
t^n - 1 &= (t^{n-1} + t^{n-2} + \cdots + t + 1)(t-1) \Rightarrow \\
\lim_{n\to\infty} t^n - 1 &= -1 = (1 + t + t^2 + t^3 + t^4 \cdots)(t-1) \quad \text{if} \quad |t| < 1 \Rightarrow \\
\frac{1}{1-t} &= 1 + t + t^2 + t^3 + t^4 + \cdots \Rightarrow \\
\frac{1}{1+t} &= 1 - t + t^2 - t^3 + t^4 - \cdots .
\end{aligned}
$$

Therefore, $\frac{1}{1+t} = 1 - t + t^2 - t^3 + \cdots$ is integrated term by term to obtain that

$$
\begin{aligned}
\ln(1+z) &= z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{z^k}{k} \Rightarrow \\
\ln(1-z) &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(-z)^k}{k}.
\end{aligned}
$$

It then follows that

$$
\begin{aligned}
\ln\left(\left(1 - \frac{\lambda}{n}\right)^n\right) &= n \ln\left(1 - \frac{\lambda}{n}\right) = n\left(-\frac{\lambda}{n} + \frac{1}{2}\frac{\lambda^2}{n^2} - \frac{1}{3}\frac{\lambda^3}{n^3} + \cdots\right) \Rightarrow \\
\lim_{n\to\infty} \ln\left(\left(1 - \frac{\lambda}{n}\right)^n\right) &= \lim_{n\to\infty} \left(-\lambda + \frac{1}{2}\frac{\lambda^2}{n} - \frac{1}{3}\frac{\lambda^3}{n^2} + \cdots\right) = -\lambda \Rightarrow
\end{aligned}
$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \exp(-\lambda).$$

Collectively, the above leads to

$$\lim_{n\to\infty}\left(\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k}\right) = \exp(-\lambda)\frac{\lambda^k}{k!}. \quad \blacksquare$$

A.4 Results on integration

Integration by substitution is a method for solving an integral, where some function $f(u)$ (which is intractable) is replaced by $f(g(x))$ [6, p. 377].

Lemma A.4.1 *Suppose that g is continuously differentiable*[3] *on an interval $[a,b]$ and that f is continuous on the interval $g([a,b])$. Then*

$$\int_a^b f(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} f(u)\,\mathrm{d}u,$$

where $u = g(x)$ and $\frac{\mathrm{d}u}{\mathrm{d}x} = g'(x)$, which in turn means that $g'(x)\,\mathrm{d}x = \mathrm{d}u$.

Integration by parts is used to transform the integral of a product of functions such that it is easier to find the integral [6, p. 522].

Lemma A.4.2 (Integration by parts) *Let $u(x), v(x) : [a,b] \to \mathbb{R}$ be two continuously differentiable functions with derivatives $u'(x), v'(x)$. Then*

$$\begin{aligned}\int_a^b u(x)v'(x)\,\mathrm{d}x &= [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)\,\mathrm{d}x \\ &= u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)\,\mathrm{d}x.\end{aligned}$$

One can repeatedly use integration by parts, which requires repeatedly finding the derivatives and antiderivatives of $u(x)$ and $v(x)$, respectively. Let $u^{(i)}(x)$ denote the i'th derivative of $u(x)$. Using integration by parts of an indefinite integral three times leads to

$$\begin{aligned}\int u^{(0)}(x)v^{(3)}(x)\,\mathrm{d}x &= u^{(0)}(x)v^{(2)}(x) - u^{(1)}(x)v^{(1)}(x) \\ &\quad + u^{(2)}(x)v^{(0)}(x) - \int u^{(3)}(x)v^{(0)}(x)\,\mathrm{d}x.\end{aligned}$$

[3] A function is continuously differentiable if its derivative exists and is also a continuous function.

More generally, repeated integration by parts of an indefinite integral k times can be written as

$$\begin{aligned}\int u^{(0)}(x)v^{(k)}(x)\,\mathrm{d}x &= u^{(0)}(x)v^{(k-1)}(x) - u^{(1)}(x)v^{(k-2)}(x) + u^{(2)}(x)v^{(k-3)}(x) \\ &\quad - u^{(3)}(x)v^{(k-4)}(x) + \cdots + (-1)^{k-1}u^{(k-1)}(x)v^{(0)}(x) \\ &\quad + (-1)^k \int u^{(k)}(x)v^{(0)}(x)\,\mathrm{d}x \\ &= \sum_{i=0}^{k-1}(-1)^i u^{(i)}(x)v^{(k-1-i)}(x) + (-1)^k \int u^{(k)}(x)v^{(0)}(x)\,\mathrm{d}x.\end{aligned}$$

Repeated integration by parts of a definite integral is similar to that of an indefinite integral except that each of the terms in the sum is replaced by $[u^{(i)}(x)v^{(k-1-i)}(x)]_a^b$, i.e.

$$\begin{aligned}\int_a^b u^{(0)}(x)v^{(k)}(x)\,\mathrm{d}x1 &= \sum_{i=0}^{k-1}(-1)^i \left[u^{(i)}(x)v^{(k-1-i)}(x)\right]_a^b + (-1)^k \int_a^b u^{(k)}(x)v^{(0)}(x)\,\mathrm{d}x \\ &= \sum_{i=0}^{k-1}(-1)^i \left(u^{(i)}(b)v^{(k-1-i)}(b) - u^{(i)}(a)v^{(k-1-i)}(a)\right) \qquad \text{(A.5)} \\ &\quad + (-1)^k \int_a^b u^{(k)}(x)v^{(0)}(x)\,\mathrm{d}x.\end{aligned}$$

A.5 Confidence intervals

Confidence intervals of a mean are used in Sections 3.3.1 and 4.1.1, and the underlying theory is described in this section. A confidence interval is firstly defined in Definition A.5.1 [5, p. 343].

Definition A.5.1 *Let n realizations $x_1, \ldots, x_n$ of random variables $X_1, \ldots, X_n$, a parameter of interest θ, and a probability γ between 0 and 1 be given. If for every value of θ there exist sample statistics*[4] *$L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that*

$$P(L_n < \theta < U_n) = \gamma$$

then the interval (l_n, u_n) with $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$ is called a $100 \cdot \gamma\%$ confidence interval for θ with confidence level γ.

Therefore, with a confidence interval one can be confident (at a given confidence level) that the true value of a parameter is inside this interval.

[4] A sample statistic can e.g. be the sample mean or variance [5, p. 254].

In this thesis, the data being analyzed with confident intervals are normally distributed, where the parameter of interest is the mean μ (e.g. the mean absolute errors in Section 4.1.1). Finding the confidence interval requires finding the critical value z_p, which is defined as

$$P(Z \geq z_p) = p, \quad Z \sim \mathcal{N}(0,1).$$

If one e.g. desires p to be 0.025, one can look up $1 - 0.025 = 0.975$ in a table of z-scores, which are available online and in statistics textbooks, and see that this corresponds to $z_p = 1.96$. Furthermore, $z_{1-p} = -z_p$ [5, p. 345].

If $X_1, \ldots, X_n$ are random variables with each $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then the mean is $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, which in turn means that [14, p. 128]

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

If the upper and lower critical values c_u, c_l are chosen such that $P(c_l < Z < c_u) = \gamma$ for $Z \sim \mathcal{N}(0,1)$, it follows that [5, p. 346]

$$\begin{aligned}\gamma = P\left(c_l < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < c_u\right) &= P\left(c_l\frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < c_u\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\bar{X}_n - c_u\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n - c_l\frac{\sigma}{\sqrt{n}}\right).\end{aligned}$$

Therefore

$$L_n = \bar{X}_n - c_u\frac{\sigma}{\sqrt{n}}, \quad U_n = \bar{X}_n - c_l\frac{\sigma}{\sqrt{n}},$$

which means that the $100 \cdot \gamma\%$ confidence interval for μ is

$$\left(\bar{x}_n - c_u\frac{\sigma}{\sqrt{n}}, \bar{x}_n - c_l\frac{\sigma}{\sqrt{n}}\right), \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The value of $\alpha = 1 - \gamma$ is usually evenly divided between the tails such that

$$\begin{aligned}P(Z \geq c_u) = \alpha/2 = P(Z \leq c_l) \Rightarrow \\ c_u = z_{\alpha/2}, \qquad c_l = z_{1-\alpha/2} = -z_{\alpha/2}.\end{aligned}$$

The $100 \cdot \gamma\%$ confidence interval for μ with a z-score as critical value is then

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \tag{A.6}$$

However, this requires that the standard deviation σ is known. If that is not the case, the random variable

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}, \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

can be used, and its distribution only depends on n and not on μ or σ [5, p. 348]. This distribution is known as the t-distribution and is defined in Definition A.5.2 [5, p. 348].

Definition A.5.2 *If the pdf of a continuous random variable X is given by*

$$f(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in \mathbb{R}, \quad k_m = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\left(\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}\right)}, \quad m \in \mathbb{N},$$

then X is said to have a t-distribution with m degrees of freedom, which is denoted by $X \sim t(m)$.

Similarly to a critical value z_p, the critical value $t_{m,p}$ for a t-distribution satisfy that $P(T \geq t_{m,p}) = p$ with $T \sim t(m)$. The critical values of a t-distribution also satisfies that $t_{m,1-p} = -t_{m,p}$ [5, p. 349].

For n random variables $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the studentized mean

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

has a $t(n-1)$-distribution [5, p. 349]. The $100 \cdot \gamma\%$ confidence interval for μ with the critical values of a t-distribution is then (similarly to (A.6))

$$\left(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right), \tag{A.7}$$

where the critical value $t_{n-1,\alpha/2}$ can e.g. be computed from $\alpha/2$ and $n-1$ with functions that are usually available in statistical software (e.g. the `t.ppf` function from the `scipy.stats` package in Python).

A.6 Stationarity of random processes

In this section, a random process $\mathbf{X}^t = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ with $t \in \mathbb{N}$ (see also Definition 2.1.1) and the stationarity of such a process is considered, which is used in Section 4.2.

The autocorrelation function is an important concept when talking about stationarity and is defined in Definition A.6.1. Note that the term *autocorrelation* means *correlation in time* and is the correlation between two points in time of the process.

Definition A.6.1 (Autocorrelation function) *For a random process $\mathbf{X}^t$, the autocorrelation function (ACF) is defined as*

$$R_{\mathbf{X}}(t_1, t_2) = \mathbb{E}\left[\mathbf{X}_{t_1}\mathbf{X}_{t_2}\right].$$

Note that

$$R_{\mathbf{X}}(t_1, t_2) = C_{\mathbf{X}}(t_1, t_2) + \mu_1\mu_2, \quad C_{\mathbf{X}}(t_1, t_2) = \operatorname{cov}(\mathbf{X}_{t_1}, \mathbf{X}_{t_2}),$$

where $\mu_1 = \mathbb{E}[\mathbf{X}_{t_1}]$ and $\mu_2 = \mathbb{E}[\mathbf{X}_{t_2}]$, respectively. This means that the autocorrelation of $\mathbf{X}_{t_1}$ and $\mathbf{X}_{t_2}$ depends on the autocovariance $C_{\mathbf{X}}$ of these. Therefore, the autocorrelation function can e.g. be used to describe repeating patterns in a process, where the autocovariance is high.

The term weak-sense stationarity (also called wide-sense stationary) is defined in Definition A.6.2. Note that a process can also be strict-sense stationary but these conditions are too strict, which means that they are rarely fulfilled in practice, and it will therefore not be considered here.

Definition A.6.2 (Weak-sense stationarity) *A random process $\mathbf{X}^t$ is said to be weak-sense stationary (WSS) if*

1. *The expected value of $\mathbf{X}^t$ is constant, i.e. $\mathbb{E}[\mathbf{X}^t] = \mu_{\mathbf{X}}$.*
2. *The ACF $R_{\mathbf{X}}(t_1, t_2)$ only depends on the time lag $\tau = |t_1 - t_2|$, i.e. $R_{\mathbf{X}}(t_1, t_2) = R_{\mathbf{X}}(0, \tau) \triangleq R_{\mathbf{X}}(\tau)$.*

The first condition in Definition A.6.2 implies that $\mathbf{X}_1, \ldots, \mathbf{X}_t$ behaves similarly, which intuitively makes sense when talking about stationarity. The second condition implies that the autocorrelation only depends on the difference between t_1 and t_2 and not on their actual position in time.

An example of a WSS process is white noise, which is defined in Definition A.6.3 [8, p. 556].

Definition A.6.3 (White noise) *A white noise process is defined as a WSS process with zero mean, constant variance σ_W^2, and uncorrelated random variables $\mathbf{W}_1, \ldots, \mathbf{W}_t$, which is denoted by $\mathbf{W}_1, \ldots, \mathbf{W}_t \sim \operatorname{wn}(0, \sigma_W^2)$.*

Note that no particular pdf is specified in Definition A.6.3. The random variables may further be independent and identically distributed (iid), which is known as iid white noise and is denoted by $\mathbf{W}_1, \ldots, \mathbf{W}_t \overset{\text{iid}}{\sim} \operatorname{wn}(0, \sigma_{\mathbf{W}}^2)$. Furthermore, a Gaussian pdf may also be specified, which is known as Gaussian white noise and is denoted by $\mathbf{W}_1, \ldots, \mathbf{W}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{W}}^2)$ [8, p. 556].

In practice, a process is often tested for stationarity with *unit root testing*, which is shortly introduced in the following.

An AR(1) process (i.e. a process containing one autoregressive term) can be written as

$$\mathbf{X}_t = \phi\mathbf{X}_{t-1} + \mathbf{W}_t, \tag{A.8}$$

where $\mathbf{W}_1, \ldots, \mathbf{W}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and ϕ is the autoregressive parameter. If $\phi = 1$, $\mathbf{X}_t$ will be equal to its own past at time $t-1$ plus some noise, which is called a random walk and which is not a stationary process. Therefore, it is generally required for an AR(1) process that $|\phi| < 1$ in order for it to be stationary. Examples of AR(1) processes with $\phi = 0.7$ and $\phi = 1.0$, respectively, are shown in Figure A.1.

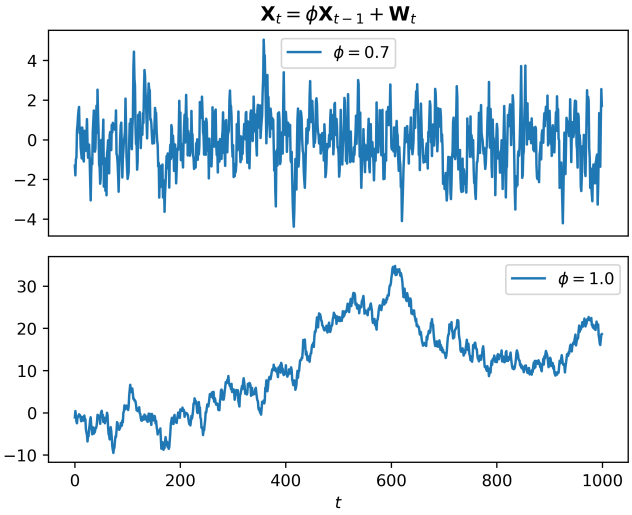


Figure A.1: Examples of AR(1) processes with $\phi = 0.7$ and $\phi = 1.0$.

Figure A.1 clearly shows that the process is nonstationary when $\phi = 1$ compared to when $\phi = 0.7$.

Random processes can be tested statistically for stationarity with a so-called unit root test. One unit root test is the Dickey-Fuller (DF) statistic, which tests whether $\phi = 1$ or $|\phi| < 1$ in (A.8), i.e. [17, p. 250]

$$H_0 : \phi = 1 \qquad \text{versus} \qquad H_1 : |\phi| < 1.$$

If the p-value e.g. is less than 0.05, this means that one can be 95% confident that the random process is stationary.

An AR(p) process with parameters $\phi_1, \ldots, \phi_p$ can more generally be written as [17, p. 76]

$$\mathbf{X}_t = \sum_{i=1}^{p} \phi_i \mathbf{X}_{t-i} + \mathbf{W}_t \Rightarrow$$
$$\mathbf{X}_t - \sum_{i=1}^{p} \phi_i \mathbf{X}_{t-i} = \mathbf{W}_t \Rightarrow$$

$$\phi(\mathrm{B})\mathbf{X}_t = \mathbf{W}_t, \quad \phi(\mathrm{B}) = 1 - \phi_1\mathrm{B} - \phi_2\mathrm{B}^2 - \cdots - \phi_p\mathrm{B}^p, \tag{A.9}$$

where B is the back-shift operator for which $\mathrm{B}^k\mathbf{X}_t = \mathbf{X}_{t-k}$. The polynomial $\phi(\mathrm{B})$ of order p in (A.9) has p roots (some of which may be complex), and the condition for stationarity of (A.9) is that neither of these roots are on the unit circle, i.e. $\{z \in \mathbb{C} : |z| \neq 1\}$, which can be tested with the augmented Dickey-Fuller (ADF) test [17, p. 252][5]. In this thesis, the random processes are tested for stationarity in Python with the function `adfuller` from the package `statsmodels.tsa.stattools`.

[5] Further descriptions of the described unit root tests and how to obtain the test statistics are out of scope for this thesis, and the reader is referred to the referenced literature.

B | Additional graphs

Figure B.1 shows a histogram of the number of samples in each session in the EEG data before being truncated as described in Section 4.3.

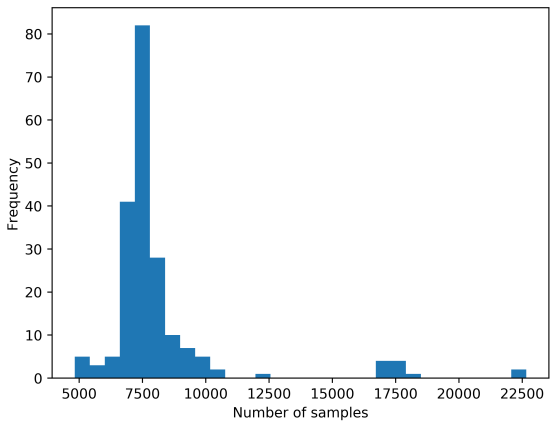


Figure B.1: Histogram of the length of the sessions in the EEG data described in Section 4.3.

The errors of the tests with different values of k, d, and n from Section 4.1.1 are tested for normality with the Shapiro-Wilk test, where a p-value above 0.05 means that the errors can be assumed to be normally distributed [13]. Further descriptions of this test is out of scope for this thesis.

Figures B.2-B.3 show scatterplots of the p-values from the Shapiro-Wilk test of the errors from the tests of different values of d, k, and n from Section 4.1.1. The red line in each figure indicates the p-value of 0.05.
The analyses in Figures B.2-B.3 show that 88.33% or more of the errors can be assumed to be normally distributed.

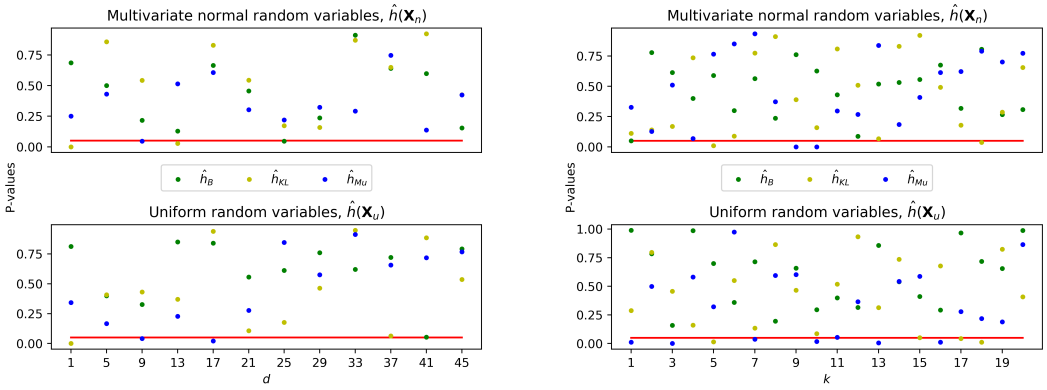


Figure B.2: P-values for Shapiro-Wilk tests of the errors of the estimates with different values of d (left) and k (right). The number of p-values below 0.05 is 7 (9.72%) and 14 (11.67%) for the tests of d and k, respectively.

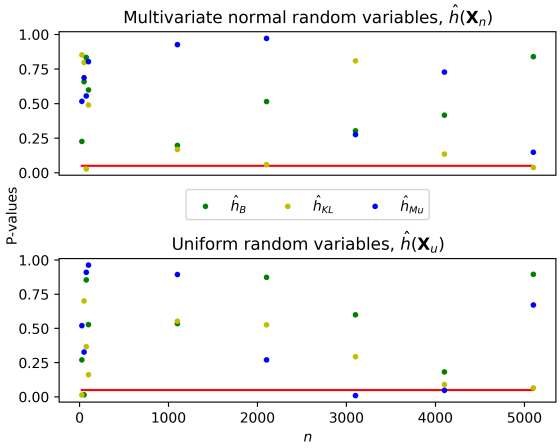


Figure B.3: P-values for Shapiro-Wilk tests of the errors of the estimates with different values of n. The number of p-values below 0.05 is 6 (11.11%).