AALBORG UNIVERSITY

STUDENT REPORT

Master's Thesis
Mathematical Engineering

Dependency Analysis of Electroencephalography Signals

A Theoretical and Data Driven Approach to Quantifying Dependencies in Multivariate Signals

Authors:
Alexander Djupnes Fuglkjær
Frederik Appel Vardinghus-Nielsen
Magnus Berg Ladefoged

Supervisors:
Christophe Biscio
Jan Østergaard
Payam Baboukani

2nd June 2023

Dept. of Mathematical Sciences
Skjernvej 4A
DK-9220 Aalborg Ø
http://math.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Dependency Analysis of Electroencephalography Signals

Theme:
A Theoretical and Data Driven Approach to Quantifying Dependencies in Multivariate Signals

Project Period:
Spring Semester 2021

Project Group:
MATTEK10 grp 4.105b

Authors:
Alexander Djupnes Fuglkjær
Frederik Appel Vardinghus-Nielsen
Magnus Berg Ladefoged

Supervisors:
Christophe Biscio
Jan Østergaard
Payam Baboukani

Copies: 3

Numbered Pages: 145

Date of Completion:
2nd June 2023

Abstract - English

In this project, information theory, graph theory, as well as the omega complexity is presented in order to analyse dependencies in EEG signals. An analysis of the omega complexity is performed, and a generalised omega complexity is introduced to combat some of the presented deficiencies. The methods are tested on coupled Rössler systems and multivariate autoregressive processes as these have proven to be comparable with EEG signals in their behaviour. Initially an EEG data set obtained from a subject exposed to a high and low SNR environment is analysed, although no significant changes between the two are found. Next, the presented methods are applied to an iEEG data set on a subject with epilepsy, resulting in significant changes between dependencies in the EEG signals prior to a seizure and during a seizure. Hence the methods introduced are to some degree able to capture changes in dependencies in EEG signals.

Abstract - Dansk

I dette projekt præsenteres informationsteori, grafteori samt omegakompleksitet med henblik på at udføre en analyse af afhængigheder i EEG signaler. Der udføres en analyse af omegakompleksiteten, og der introduceres en generaliseret udgave af omegakompleksitet for at forbedre nogle af de præsenterede mangler. Metoderne testes på koblede Rösslersystemer og multivariate autoregressive processer, da disse har vist sig i nogen grad at have EEG-lignende opførsel. Først analyseres et EEG-datasæt, der er indsamlet fra en person udsat for et høj- og lav-SNR-miljø, dog observeres der ingen signifikante ændringer mellem de to. Herefter anvendes de introducerede metoder på et iEEG-datasæt fra en person med epilepsi, hvilket viser signifikante ændringer i afhængigheder i EEG-signalerne før og under et anfald. Dermed er metoderne i nogen grad i stand til at fange ændringer i afhængigheder i EEG-signaler.

Preface

This project is written in the period 01/09/22 to 02/05/23 by the group MATTEK10 4.105b attending fourth semester of the Master's program of Mathematical Engineering at Aalborg University.

The results throughout the project are calculated using Python v3.10.7 with the packages Numpy v1.23.5, Scipy v1.9.3, Pandas v1.5.2 and scikit-learn v1.1.3. The figures were generated using matplotlib v3.6.2. In addition, Matlab 2023b utilizing the ITS package v2.1 Faes [2019] was used.

The group would like to thank Christophe Biscio, Jan Østergaard and Payam Baboukani for supervision throughout the process of writing this project.

Secondly the group would like to thank the other two MATTEK10 groups for more or less productive discussions throughout the year, as well as way too hot group rooms with hazardously low levels of oxygen concentration.

Accompanying scripts to the project can be found attached to the project as `scripts.zip`.

Alexander Djupnes Fuglkjær

Frederik Appel Vardinghus-Nielsen

Magnus Berg Ladefoged

Contents

1 | Problem Analysis

1.1 Introduction

Hearing loss is a growing problem worldwide. In 2023 more than 5% of people suffer from a disabling hearing loss, characterised as a hearing loss of more than 35dB, and it is estimated that by 2050 this will effect more than 10% [World Health Organization, 2023]. To improve the hearing of people suffering from hearing loss, an option could be a hearing aid [University of California San Fransisco, 2023]. One of the worlds leading manufacturers of hearing aids have deployed hearing aids which introduce technology based upon electroencephalography signals [Santurette et al., 2020]. They have also claimed that the BrainHearing™ [Oticon, 2023] technology of their hearing aids causes a decrease in a term coined listening effort [Nielsen and Ng, 2022]. In [Baboukani et al., 2022], it is stated that listening effort can be examined in various ways, but that estimating listening effort through electroencephalography measurements has gained popularity.

1.2 Electroencephalogram

Electroencephalography (EEG) is a non-intrusive measurement of the brain's activity [Biasiucci et al., 2019]. It is carried out by placing a cap on the subject, in which a number of electrodes is placed. These electrodes then monitor the voltage potentials that arise as a result of the neurons firing in the brain. As the electrodes are not placed directly on the brain, the problem of volume conduction arises. This means that all electrodes to some degree capture a mix of potentials from all over the brain. Because of this EEGs unfortunately have very poor spatial resolution. Furthermore EEGs are often contaminated by a number artifacts, for example muscle movements [Muthukumaraswamy, 2013], sweat, eye movements, and more [Britton et al., 2016] . However, EEGs have a very high temporal resolution, which allows for detection of rapid changes in the activity of the subject's brain.

To get an understanding of how the electrodes are placed on the scalp of the subject, the placements of the electrodes for a 64 electrode EEG measurement are shown in Figure 1.1.

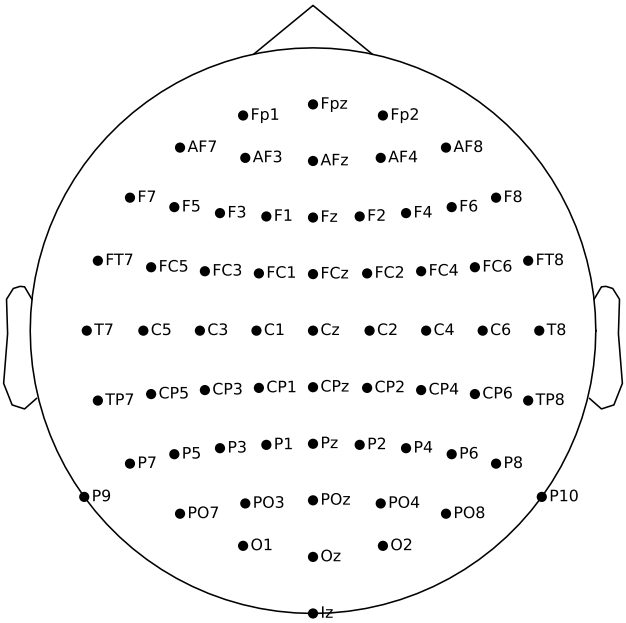


Figure 1.1: EEG electrode placement.

1.2.1 EEG signals

The output of an EEG measurement is a series of signals, with the number of signals determined by the number of electrodes used. An example of a subset of EEG signals is seen in Figure 1.2. EEG signals have been divided into specific bands in the

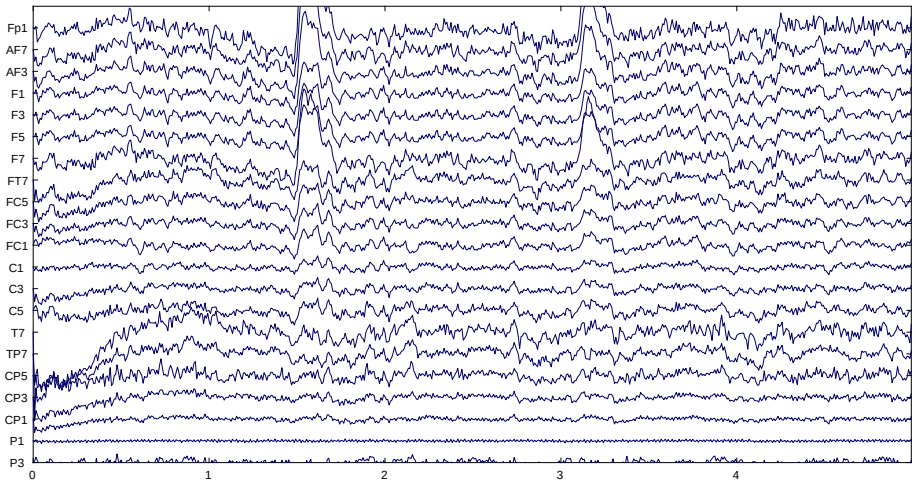


Figure 1.2: Example of five seconds of an EEG from 15 electrodes.

frequency domain namely the delta, theta, alpha, beta, and gamma bands [Ameera et al., 2018]. These frequency bands allow a more specific analysis of the EEG signals [Baboukani et al., 2022] and the bands can be seen in Table 1.1.

Band	Freq.
Delta	$1 - 4$Hz
Theta	$4 - 8$Hz
Alpha	$8 - 13$Hz
Beta	$13 - 30$Hz
Gamma	> 30Hz

Table 1.1: Named frequency bands of EEG signals.

1.3 Listening Effort

Listening effort can be defined as "the mental exertion required to attend to, and understand, an auditory message" [Miles et al., 2017]. Listening effort has become a major topic of research in the auditory field as it has been shown that high listening effort is linked to mental and physical fatigue, and can in part have consequences on social and work life [Nielsen and Ng, 2022]. It is therefore a very important research topic for manufacturers of hearing aids, as these seek to reduce listening effort among those using their devices.

Studies have shown that it is not only intelligibility which is affected by listening effort but that high listening effort can have a negative impact on language processing and even on memory [Peele, 2017].

Information about listening effort can be obtained by a questionnaire where the subjects lists their perceived listening effort on some set scale, but this can introduce numerous errors. It has also been shown that the mentality of the participants can have a great impact on the results of such tests, which complicates variable control McGarrigle et al. [2014]. Another way of obtaining information about listening effort could be through EEG signals.

In Baboukani et al. [2022] it is stated that there are a number of signal processing tools, for example quantification of statistical dependencies, that can be applied to EEG signals to find correlates of listening effort and multiple studies have shown a potential of EEG signals to capture changes that occur when a subject listens to a source which is more difficult to listen to, for example with low SNR [Ala et al., 2022].

1.4 General Dependency Analyses of EEG Signals

The task of examining changes in EEG signals of a subject under different conditions is not only applicable in auditory research. For example in Mensen et al. [2017] differentiation analysis is used to quantify neurophysiological differentiation through EEG signals based on a subject watching pictures of different personal importance. In Kramer et al. [2008] the authors examined changes in EEG signals of a subject before and during an epileptic seizure and found a decrease in connectivity of the

brain immediately before and during seizure. Hence the task of monitoring changes in EEG signals through dependency analyses is a relevant topic not only in terms of listening effort but as a tool for explaining neural activity.

Different mathematical tools have been used in order to quantify changes in EEG signals. In [Wackermann, 1996] an approach to analysing EEG signals called the omega complexity was introduced. The results in the paper showed that the omega complexity could reflect physiological states of the human brain. In [Baboukani et al., 2022] this tool was used to conclude that phase synchrony in EEG signals change under low and high SNR conditions when a noise reduction scheme is activated. In [Xefteris et al., 2022] graph theory in conjunction with mutual information was used to analyse both local and global features of EEG signal which where then used for recognition of emotions in the subject.

1.5 Problem Statement

Based on the problem analysis, a relevant contribution to the research regarding listening effort, could be to examine dependencies of EEG signals in a more general framework. To this end it seems fitting to explore the possibility of using information theory in conjunction with the omega complexity to quantify said dependencies. Furthermore it is hypothesized that relevant information about listening effort is contained within a subset of the EEG signals and hence graph theory with a data driven approach for selecting subsets of electrodes for analysis is explored.

These considerations in combination lead to the following problem statement:

How can information theory, graph theory and the omega complexity be used to analyse the activity of the human brain under different conditions through quantification of dependencies in EEG signals?

2 | The Hilbert Transform

This chapter introduces the concept of an analytic signal, showcases some of its properties and relates it to the Hilbert transform. The chapter is based upon Liu [2012], Marple [1999] and Kschischang [2015].

2.1 Analytic Signal

A real-valued signal $s_r(t)$ has a symmetric Fourier magnitude spectrum and hence contains negative frequencies. The signal can, however, be converted to an analytic signal $s_a(t)$ which has no negative frequency components.

Definition 2.1 (Analytic Signal)
A signal with no negative frequency components is called an analytic signal.

In the following section, a method for obtaining an analytical representation of a real valued signal will be presented.

2.2 The Hilbert Transform

The analytic representation of a real-valued signal can be related to the real signal through the Hilbert transform, which relies on the Cauchy Principal Value.

Definition 2.2 (Cauchy Principal Value (CPV))
Let $f : \mathbb{R} \to \mathbb{R}$ be a function which has a singularity at either $b \in \mathbb{R}$ or at $\pm\infty$. Consider the integral

$$\int_a^c f(x)\ \mathrm{d}x,$$

where $a < b < c$ and $a, c \in (-\infty; \infty)$. The CPV, denoted $P\{\cdot\}$, for a singularity at b is defined as

$$P\left\{\int_a^c f(x)\ \mathrm{d}x\right\} = \lim_{\epsilon \to 0^+}\left[\int_a^{b-\epsilon} f(x)\ \mathrm{d}x + \int_{b+\epsilon}^c f(x)\ \mathrm{d}x\right], \quad \text{for} \quad a < b < c,$$

provided that the limit exists. For a singularity at $\pm\infty$, the CPV is defined as

$$P\left\{\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x\right\} = \lim_{\epsilon\to\infty}\int_{-\epsilon}^{\epsilon} f(x)\ \mathrm{d}x,$$

provided that the limit exists.

The CPV seeks to assign a value to an improper integral which otherwise would not converge toward a finite value.

Example 2.3 (Cauchy Principal Value)
Consider the function $f(x) = 1/x$ with a singularity at $x = 0$ and the integral

$$\int_{-\infty}^{\infty} \frac{1}{x}\ \mathrm{d}x,$$

which is not defined. Now consider the CPV

$$\begin{aligned} P\left\{\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x\right\} &= \lim_{\epsilon\to 0}\left(\int_{-1/\epsilon}^{-\epsilon}\frac{1}{x}\ \mathrm{d}x + \int_{\epsilon}^{1/\epsilon}\frac{1}{x}\ \mathrm{d}x\right) \\ &= \ln(\epsilon) - \ln(1/\epsilon) + \ln(1/\epsilon) - \ln(\epsilon) \\ &= 0, \end{aligned}$$

which converges toward zero because of the symmetric integration around $x = 0$.

The Hilbert transform utilises the CPV in order to assign a value to a convolution involving a function with a singularity.

Definition 2.4 (The Hilbert Transform)
Let $u(t)$ be a continuous-time real-valued function. The continuous Hilbert transform of $u(t)$ is defined as

$$\mathcal{H}(u)(t) = \frac{1}{\pi}P\left\{\int_{-\infty}^{\infty}\frac{u(\tau)}{t-\tau}\ \mathrm{d}\tau\right\} = \frac{1}{\pi t} * u(t),$$

provided that the integral exists as a CPV.

Let $u(n)$ be a discrete-time real-valued function. The discrete Hilbert transform of $u(n)$ is defined as

$$\mathcal{H}(u)(n) = \begin{cases} \dfrac{2}{\pi}\displaystyle\sum_{k\ \text{odd}}\frac{u(k)]}{n-k}, & \text{for } n \text{ even} \\ \dfrac{2}{\pi}\displaystyle\sum_{k\ \text{even}}\frac{u(k)}{n-k}, & \text{for } n \text{ odd} \end{cases}.$$

The Hilbert transform is in the next section related to the the Fourier transform.

2.2.1 Relation to the Fourier Transform

The Hilbert transform is related to the Fourier transform through

$$\mathcal{F}(\mathcal{H}(u))(\omega) = -i\,\mathrm{sign}(\omega)\mathcal{F}(u)(\omega),$$

where $\mathcal{F}(u)(t)$ is the Fourier transform of $u(t)$. The Hilbert transform is furthermore an anti-involution – that is

$$\mathcal{H}(\mathcal{H}(u))(t) = -u(t) \quad \Leftrightarrow \quad \mathcal{H}^{-1} = -\mathcal{H}.$$

The connection between an analytic representation of a signal and the Hilbert transform of said signal is demonstrated by letting the Fourier transform of a function $u(t)$ fulfill the properties of an analytic signal such that

$$\begin{aligned}\mathcal{F}(u)(\omega) &= \begin{cases} 2\mathcal{F}(u)(\omega), & \text{for } \omega > 0 \\ \mathcal{F}(u)(\omega), & \text{for } \omega = 0 \\ 0, & \text{for } \omega < 0 \end{cases} \\ &= \mathcal{F}(u)(\omega) + \mathrm{sign}(\omega)\mathcal{F}(u)(\omega)\end{aligned}$$

which preserves the power of $\mathcal{F}(u)(\omega)$. By inverse Fourier transforming:

$$\begin{aligned}\mathcal{F}^{-1}\{\mathcal{F}(u)(\omega)\} &= \mathcal{F}^{-1}\{\mathcal{F}(u)(\omega) + \mathrm{sign}(\omega)\mathcal{F}(u)(\omega)\} \\ &= \mathcal{F}^{-1}\{\mathcal{F}(u)(\omega)\} + \mathcal{F}^{-1}\{\mathrm{sign}(\omega)\} * \mathcal{F}^{-1}\{\mathcal{F}(u)(\omega)\} \\ &= u(t) + \frac{i}{\pi t} * u(t) \\ &= u(t) + i\mathcal{H}(u)(t).\end{aligned}$$

This shows the relationship between an analytic representation and the Hilbert transform.

This relationship is illustrated in Figure 2.1 showing the real signal

$$s_r(t) = (1 + \sin(\omega_1 2\pi t))\sin(\omega_2 2\pi t), \tag{2.1}$$

where $\omega_1 = 1$ and $\omega_2 = 10$ along with the Hilbert transform $s_i(t) = \mathcal{H}(s_r)(t)$. Figure 2.2 shows the effect of applying the Fourier transform on $s_r(t)$ and $s_a(t) = s_r(t) + is_i(t)$.

2.3 Instantaneous Phase and -Amplitude

Rewriting the analytic signal into polar representation allows further analysis of the signal:

$$\begin{aligned}s_a(t) &= s_m(t)(\cos(\phi(t)) + i\sin(\phi(t))) \\ &= s_m(t)e^{i\phi(t)}.\end{aligned} \tag{2.2}$$

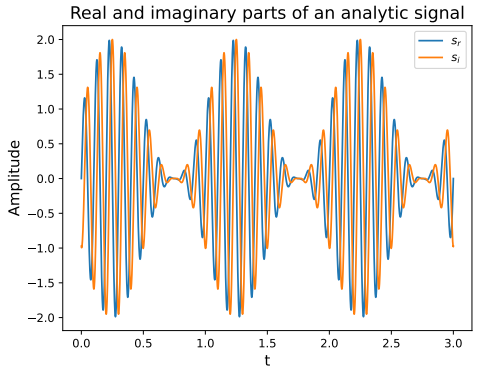


Figure 2.1: Real signal s_r along with its Hilbert transform s_i.

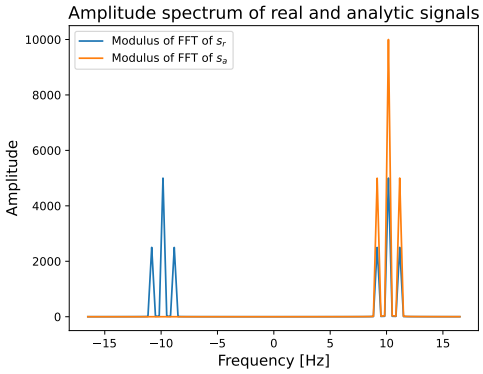


Figure 2.2: Magnitude spectra of the real signal $s_r(t)$ and analytic signal $s_a(t)$.

The instantaneous amplitude and instantaneous phase of a signal can be defined based upon the polar representation in Equation (2.2).

Definition 2.5 (Instantaneous Amplitude and Phase)
Let $s_a(t) = s_r(t) + is_i(t) = s_m(t)e^{i\phi_s(t)}$ be an analytic signal. Then

$$s_m(t) = |s_a(t)|$$

is called the instantaneous amplitude of $s_a(t)$ and

$$s_\phi(t) = \arg(s_a(t))$$

is called the instantaneous phase of $s_a(t)$.

The instantaneous amplitude can be utilised to construct an envelope based upon the analytic representation. Since $|s_a(t)| = |s_r(t) + s_i(t)| \geq |s_r(t)|$ it follows that $|s_a(t)|$ and $-|s_a(t)|$ are upper and lower bounds, respectively, of $s_r(t)$ – this particular envelope will from here on be referred to as the instantaneous amplitude. This and instantaneous phase of the signal in Equation (2.1) can be seen in Figure 2.3.

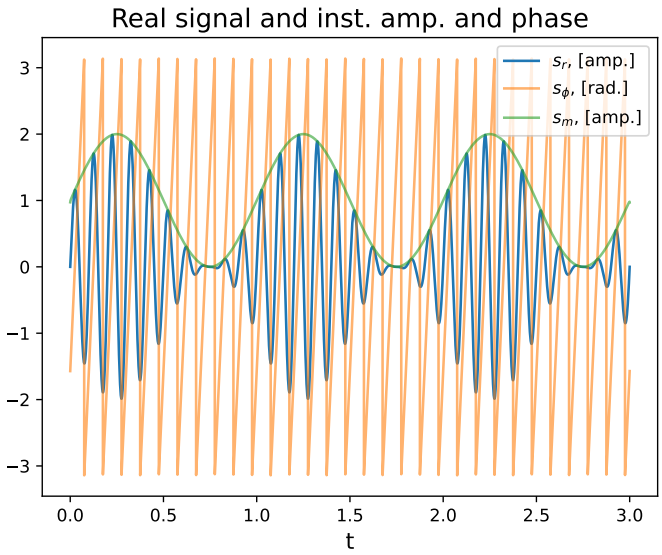


Figure 2.3: Real signal in Equation (2.1) along with instantaneous amplitude and phase from the analytic representation.

2.3.1 Noisy Instantaneous Amplitude and Instantaneous Phase

The Hilbert transform performs a phase shift of a signal and since it is an integral transform it is linear. This causes the Hilbert transform to preserve noise in the signal:

$$\mathcal{H}(u+\varepsilon)(t) = \mathcal{H}(u)(t) + \mathcal{H}(\varepsilon)(t)$$

This causes the analytic representation $s_a(t)$ of $s_r(t) = u(t) + \varepsilon(t)$ to be noisy:

$$\begin{aligned} s_a(t) &= u(t) + \varepsilon(t) + i(\mathcal{H}(u)(t) + \mathcal{H}(\varepsilon)(t)) \\ &= u_a(t) + \varepsilon_a(t) \\ &= u_m(t)e^{iu_\phi(t)} + \varepsilon_m(t)e^{i\varepsilon_\phi(t)}. \end{aligned}$$

This entails that the instantaneous amplitude and instantaneous phase of a noisy signal are noisy too. In Figure 2.4 is seen the signal in Equation (2.1) with added noise $\epsilon \sim \mathcal{N}(0, 0.005)$. It is clearly seen how noise affects the instantaneous envelope and phase – when the noise free signal has low instantaneous amplitude compared to the noise, the noise dominates the analytic signal and the envelope and instantaneous phase appear as noise.

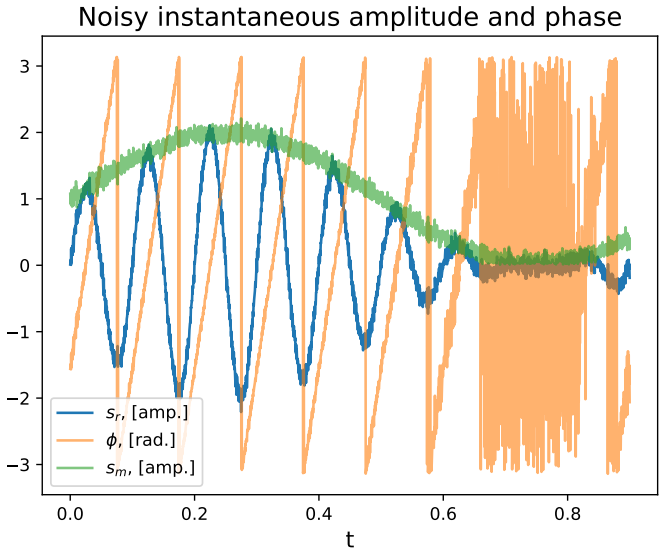


Figure 2.4: Real signal in Equation (2.1) with added noise. The instantaneous envelope and phase are influenced by noise.

3 | Circular Distributions

The phase information about a signal expressed in radians is inherently supported only on the interval $[0; 2\pi)$ (equal to any translated real interval of length 2π) with no distinction between the ends of the chosen interval – that is, the interval is considered circular and can be visualised as the unit circle with values expressed as radians. It is therefore fitting to model phase information of a signal as a stochastic variable with a circular distribution and the accompanying tools for analysis of such distributions. This chapter is based upon NCSS [2023] and Fisher and Lee [1983].

Definition 3.1 (Circular Distribution)
A probability distribution of a stochastic variable whose sample space represents angles is called a circular distribution.

The nature of circular distributions requires a different approach to calculating moments and similarity measures.

Definition 3.2 (Circular Expected Value and Circular Variance)
Let Φ be a circularly distributed stochastic variable assuming values in $[0; 2\pi)$ and with probability density function $\rho(\phi)$. The circular expected value (CEV), μ_Φ of Φ is then defined as

$$\mu_\Phi = \arg\left(E[e^{i\Phi}]\right) = \arg\left(\int_0^{2\pi} \rho(\phi)e^{i\phi}\ \mathrm{d}\phi\right),$$

and the circular variance (CV) σ_Φ as

$$\sigma_\Phi = 1 - \left|\int_0^{2\pi} \rho(\phi)e^{i\phi}\ \mathrm{d}\phi\right|.$$

Notice, that the CEV in Definition 3.2 applies the function $f(\phi) = e^{i\phi}$ before applying the expected value operator and the angle operator. This agrees with the characteristic of the distribution of Φ where the endpoints are interpreted as next to each other while it compromises some of the properties of expected values, for example linearity, non-degeneracy and multiplicativity in case of independence. The

CV quantifies the dispersion of direction by comparing length of the complex numbers $e^{i\phi}$ weighted with $\rho(\phi)$ with the maximum length 1. The variance is therefore limited to the interval $[0;1]$. The sample circular mean $\overline{\phi}$ and variance $\overline{\sigma}_\Phi$ of a sequence $\{\phi_k\}_{k=1}^N$ of samples drawn from a circular distribution can be calculated as

$$\overline{\phi} = \arg\left(\sum_{k=1}^{N} e^{i\phi_k}\right) \quad \text{and} \quad \overline{\sigma}_\Phi = 1 - \frac{1}{N}\left|\sum_{k=1}^{N} e^{i\phi_k}\right|.$$

3.1 Circular Correlation Coefficient

In order to quantify correlation between two circularly distributed variables a circular correlation coefficient (CCC) is defined.

Definition 3.3 (Circular Correlation Coefficient (CCC))
Let Φ and Θ be circular stochastic variables. Let $f(\phi,\theta)$ be their joint probability distribution, $0 \leq \phi < 2\pi$, $0 \leq \theta < 2\pi$ and μ_Φ and μ_Θ be the expected values of Φ and Θ, respectively. The circular correlation coefficient ρ_c of Φ and Θ is then defined as

$$\rho_c(\Phi,\Theta) = \frac{E[\sin(\Phi - \mu_\Phi)\sin(\Theta - \mu_\Theta)]}{\sqrt{E[\sin^2(\Phi - \mu_\Phi)]E[\sin^2(\Theta - \mu_\Theta)]}}.$$

Theorem 3.4 (Properties of the CCC)
The CCC has the following properties:

- it does not depend on the zero direction used as reference,
- it is symmetric, that is $\rho_c(\Phi,\Theta) = \rho_c(\Theta,\Phi)$,
- $-1 \leq \rho_c \leq 1$,
- if Φ and Θ are independent, then $\rho_c(\Theta,\Phi) = 0$,
- if Φ and Θ have full support, then $\rho_c(\Phi,\Theta) = \pm 1$ if and only if

$$\Phi = \pm\Theta + \gamma \mod 2\pi$$

for some constant $\gamma \in [0, 2\pi)$.

The proof of Theorem 3.4 has been omitted.

For sample sequences $\{\phi_k\}_{k=1}^N$ and $\{\theta_k\}_{k=1}^N$ from Φ and Θ, respectively, the sample CCC, $\overline{\rho_c}$, is calculated as

$$\overline{\rho_c} = \frac{\sum_{i=1}^{N} \sin\left(\phi_i - \overline{\phi}\right)\sin\left(\theta_i - \overline{\theta}\right)}{\sqrt{\sum_{j=1}^{N} \sin^2(\phi_j - \overline{\phi}) \sum_{k=1}^{N} \sin^2(\theta_k - \overline{\theta})}}.$$

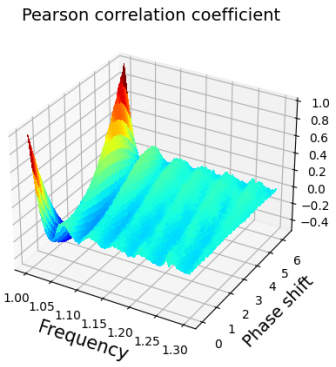


Figure 3.1: PCC between instantaneous phase of phase- and frequency shifted sines.

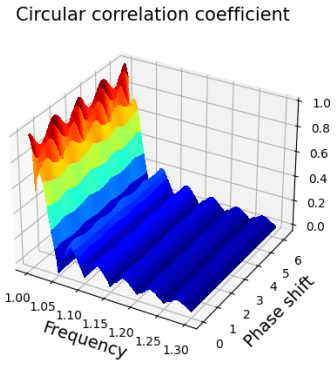


Figure 3.2: CCC between instantaneous phase of phase- and frequency shifted sines.

3.1.1 Comparison to the Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) is a well-known method of quantifying correlation between stochastic variables based upon their expected values and variances. To visualise the difference between the PCC and the CCC a set of signals $\{s_{kl}(t)\}_{k,l\in N}$, $N = \{1, \ldots, 100\}$ are generated according to

$$s_{kl}(t) = \sin\left(2\pi t\left(1 + \frac{0.3k}{100}\right) + 2\pi\frac{l}{100}\right),$$

such that $\{s_{kl}\}_{k,l\in N}$ consists of sine functions with frequencies in the interval $[1; 1.3]$ Hz and phase shifts in the interval $[0; 2\pi]$. The instantaneous phases from Definition 2.5 of these signals are correlated through the PCC and CCC to the instantaneous phase of a benchmark signal $s_b(t) = \sin(2\pi t)$ for $t \in [0; 60]$. The results can be seen in Figures 3.1 and 3.2. It is clear that the CCC is not as sensitive to phase shifts as the PCC, as evident by the difference in variability along the phase shift axis.

4 | Spectra of Hermitian Matrices

This chapter presents results regarding changes in the spectra of Hermitian matrices caused by changes in the entries of the matrices. The results become relevant in the analysis of the omega complexity in Chapter 7. This chapter is based upon Serre [2010].

Theorem 4.1 (Weyl's Inequalities)
Let $A = B + C$ and $B, C \in \mathbb{R}^{n \times n}$ be Hermitian matrices with respective eigenvalues $\lambda_i(A)$, $\lambda_i(B)$ and $\lambda_i(C)$, for $i \in \{1, \ldots, n\}$, which are ordered according to

$$\begin{aligned} \lambda_1(A) \geq \cdots \geq \lambda_n(A), \\ \lambda_1(B) \geq \cdots \geq \lambda_n(B), \\ \lambda_1(C) \geq \cdots \geq \lambda_n(C). \end{aligned}$$

Then

$$\begin{aligned} \lambda_k(A) \leq \lambda_i(B) + \lambda_j(C) \quad \text{for} \quad i + j = k + 1, \\ \lambda_k(A) \geq \lambda_i(B) + \lambda_j(C) \quad \text{for} \quad i + j = k + n. \end{aligned}$$

Theorem 4.1 establishes an ordered relationship between the eigenvalues of two Hermitian matrices and their sum showing how the spectrum of $A = B + C$ is dependent on the spectra of B and C.

Definition 4.2 (Spectral Radius)
The spectral radius $\beta(\cdot)$ of a matrix $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$ is defined as

$$\beta(A) = \max(|\lambda_1(A)|, \ldots, |\lambda_n(A)|).$$

Spectral radius is used in the following inequalities in order to establish bounds on the changes of spectra of Hermitian matrices.

Corollary 4.3 (Change of the Spectrum of a Hermitian Matrix)
Let A, B and C and their eigenvalues be defined as in Theorem 4.1. Then

$$|\lambda_i(A) - \lambda_i(B)| \leq \beta(C) \quad \text{for} \quad i \in \{1, \ldots, n\}.$$

Corollary 4.3 states that the change of the spectrum of a Hermitian matrix B when a matrix C is added to it is bounded by the spectral radius of C. The change in eigenvalues is therefore bounded.

Lemma 4.4
Let $\|\cdot\|$ be a sub-multiplicative matrix norm on $\mathbb{R}^{n\times n}$ and $A \in \mathbb{R}^{n\times n}$. Then

$$\beta(A) \leq \|A\|.$$

Proof: Let $\lambda = \max(\{\lambda_1(A), \ldots, \lambda_n(A)\})$ and $x \in \mathbb{R}^n \backslash \{0\}$ the corresponding eigenvector to λ. Define $X = [x|\cdots|x] \in \mathbb{R}^{n\times n}$ as a matrix consisting of copies of x such that

$$AX = \lambda X.$$

It then follows by the sub-multiplicativity of $\|\cdot\|$ that

$$|\lambda|\|X\| = \|\lambda X\| = \|AX\| \leq \|A\|\|X\| \quad \Rightarrow \quad |\lambda| \leq \|A\|.$$

■

Corollary 4.4 shows that the spectral radius is bounded by an arbitrary sub-multiplicative matrix norm which facilitates the following corollary.

Corollary 4.5
Let A, B and C and their eigenvalues be defined as in Theorem 4.1. Then

$$|\lambda_i(A) - \lambda_i(B)| \leq \|C\| \quad \text{for} \quad i \in \{1, \ldots, n\},$$

where $\|\cdot\|$ is a sub-multiplicative norm.

Proof: The proof follows from Corollary 4.3 and Lemma 4.4.

■

Corollary 4.5 establishes that the difference between eigenvalues of the Hermitian matrix $A = B + C$ and B are bounded by a norm of C.

The remainder of this chapter proves a linear relationship between the maximum eigenvalue of a certain type of matrix and changes of the entries of this matrix, which will be relevant in examining the omega complexity in Chapter 7.

Lemma 4.6 (Matrix Determinant Lemma)
Let $A \in \mathbb{R}^{n\times n}$ and $u, v \in \mathbb{R}^n$. Then

$$\det(A + uv^T) = \det(A) + v^T \text{adj}(A)u,$$

where $\text{adj}(A)$ is the adjugate of A.

Proof: Firstly, the case where $A = I$ is proved. Consider

$$\begin{bmatrix} I & 0 \\ v^T & 1 \end{bmatrix} \begin{bmatrix} I + uv^T & u \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ -v^T & 1 \end{bmatrix} = \begin{bmatrix} I & u \\ 0 & 1 + v^T u \end{bmatrix}.$$

Since all matrices on the LHS are square, it follows that the determinant of them is multiplicative. Furthermore, since the first and third matrix on the LHS are triangular, and the diagonals consist solely of ones, their determinants are 1. Finally, since the matrix on the RHS is triangular it follows that the eigenvalues are found on the diagonal. As the determinant is the product of all eigenvalues and all but one entry on the diagonal of the matrix on the RHS are 1, it follows that

$$\det\left(\begin{bmatrix} I & u \\ 0 & 1 + v^T u \end{bmatrix}\right) = 1 + v^T u.$$

Thus

$$\det(I + uv^T) = 1 + v^T u.$$

The general case is then achieved by using the above and the multiplicativity of the determinant such that

$$\begin{aligned} \det(A + uv^T) &= \det(A)\det(I + (A^{-1}u)v^T) \\ &= \det(A)(1 + v^T(A^{-1}u)). \end{aligned}$$

Furthermore since $\text{adj}(A)A = \det(A)I$ it follows that

$$\det(A)\Big(1 + v^T(A^{-1}u)\Big) = \det(A) + v^T \text{adj}(A)u.$$

■

Lemma 4.6 provides a possibly simpler way to calculate determinants of certain matrices. This is used in the following theorem.

Theorem 4.7 (Linearity of Eigenvalues)
Let $A \in \mathbb{R}^{n\times n}$ be on the form

$$A = \begin{bmatrix} k & \mu & \cdots & \mu \\ \mu & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \mu \\ \mu & \cdots & \mu & k \end{bmatrix}, \quad \mu \in [0; k], k > 0.$$

Then A has two eigenvalues λ_1 and λ_2 of multiplicity 1 and $n-1$, respectively, which are given by

$$\lambda_1 = k + \mu(n-1)$$
$$\lambda_2 = k - \mu.$$

Proof: The matrix A can be written as $I(k-\mu) + \mu e^T e$ where $e = [1, \ldots, 1]$. Hence finding its eigenvalues amounts to solving $\det(I(k-\mu-\lambda) + \mu e^T e) = 0$. By Lemma 4.6:

$$\det\Big(I(k-\mu-\lambda) + \mu e^T e\Big) = \det(I(k-\mu-\lambda)) + \mu e \ \mathrm{adj}(I(k-\mu-\lambda))e^T.$$

The equation to be solved is then

$$\det(I(k-\mu-\lambda) + \mu e \ \mathrm{adj}(I(k-\mu-\lambda))e^T) = 0.$$

Observe then since $\mathrm{adj}(A) = \det(A)A^{-1}$ that

$$\mathrm{adj}(I(k-\mu-\lambda)) = \begin{bmatrix} (k-\mu-\lambda)^{n-1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & (k-\mu-\lambda)^{n-1} \end{bmatrix},$$

which yields

$$\begin{aligned} 0 &= \det(I(k-\mu-\lambda)) + \mu e \ \mathrm{adj}(I(k-\mu-\lambda))e^T \\ &= (k-\mu-\lambda)^n + \mu \sum_{i=1}^{n} (k-\mu-\lambda)^{n-1} \\ &= (k-\mu-\lambda)^n + \mu n (k-\mu-\lambda)^{n-1} \\ &= (k-\mu-\lambda)^{n-1}(\mu n + (k-\mu-\lambda)), \end{aligned}$$

implying that

$$\lambda = k - \mu \quad \vee \quad \lambda = k + \mu(n-1).$$

■

Hence for a matrix as specified in Theorem 4.7, there is a linear relationship between changes in the off-diagonal entries and the eigenvalues.

5 | Information Theory

In this chapter concepts from information theory are introduced with the purpose quantifying similarity between multiple signals. This chapter is based upon Cover and Thomas [2016], and Rosas et al. [2019].

5.1 Shannon Entropy

Consider a discrete stochastic variable X with sample space $\mathcal{S}$ of cardinality $|\mathcal{S}|$. To uniquely specify the outcome of such a variable requires $\log_2(|\mathcal{S}|)$ bits. Given knowledge about the distribution of X, the average number of bits required to specify the outcome, called the Shannon entropy, can be calculated. The Shannon entropy quantifies the uncertainty about the outcome of a stochastic variable given its distribution and lays the foundation for more advanced information theoretic concepts. For the remainder of the project log refers to $\log_2$ unless otherwise specified.

Definition 5.1 (Shannon Entropy)
Let X be a discrete stochastic variable with sample space $\mathcal{S}$. The entropy of X is

$$H(X) = -\sum_{x \in \mathcal{S}} p(x) \log(p(x)),$$

where $p(x)$ is the probability mass function of X.

Since for all $x \in (0,1]$, $\log(x) \leq 0$ and by convention $0 \log(0) \triangleq 0$, it follows that the Shannon entropy is non-negative. The Shannon entropy will in the remainder of the project be referred to as entropy.

Note that $H(X) \leq \log(|\mathcal{S}|)$ with equality achieved only in the case of a uniformly distributed X – that is, the uniform distribution maximises entropy for a discrete variable with a certain sample space. Hence a concept of distance to uniformity is introduced.

Definition 5.2 (Negentropy)
Let X be a discrete stochastic variable with sample space $\mathcal{S}$. The negentropy of X is

$$N(X) = \log(|\mathcal{S}|) - H(X).$$

Negentropy in some sense quantifies how far X is from being uniformly distributed, and how much information about X is revealed through its distribution. Notice that the maximal entropy of X can be expressed as:

$$\log(|\mathcal{S}|) = N(X) + H(X)$$

A natural way to extend the concepts of entropy and negentropy is to consider the uncertainty about the outcome of multiple stochastic variables.

Definition 5.3 (Joint Entropy)
Let $X_1, \ldots, X_n$ be discrete stochastic variables with sample spaces $\mathcal{S}_1, \ldots, \mathcal{S}_n$, respectively. The joint entropy of $X_1, \ldots, X_n$ is

$$H(X_1, \ldots, X_n) = -\sum_{x_1 \in \mathcal{S}_1, \ldots, x_n \in \mathcal{S}_n} p(x_1, \ldots, x_n) \log(p(x_1, \ldots, x_n)),$$

Definition 5.4 (Joint Negentropy)
Let $X_1, \ldots, X_n$ be discrete stochastic variables with sample spaces $\mathcal{S}_1, \ldots, \mathcal{S}_n$, respectively. The joint negentropy of $X_1, \ldots, X_n$ is

$$N(X_1, \ldots, X_n) = \sum_{k=1}^{n} \log(|\mathcal{S}_k|) - H(X_1, \ldots, X_n).$$

As with the entropy and negentropy in Definitions 5.1 and 5.2, the following can be obtained directly from Definition 5.3:

$$\sum_{k=1}^{n} \log(|\mathcal{S}_k|) = N(X_1, \ldots, X_n) + H(X_1, \ldots, X_n).$$

Now consider that knowledge of the outcome of one stochastic variable can decrease the uncertainty about the outcome of another. This leads to the definition of conditional entropy.

Definition 5.5 (Conditional Entropy)
Let X,Y be discrete stochastic variables with sample spaces $\mathcal{S}$ and $\mathcal{T}$, respectively. The conditional entropy of X given Y is

$$H(X|Y) = -\sum_{x \in \mathcal{S}, y \in \mathcal{T}} p(x, y) \log\left(\frac{p(x, y)}{p(y)}\right).$$

Note that $H(X|Y)$ can be interpreted as the remaining uncertainty about X given knowledge about Y. Furthermore note that the uncertainty about X cannot increase with knowledge of Y. Joint and conditional entropy are related through the following theorem.

Theorem 5.6 (Chain Rule of Entropy)
Let X and Y be discrete stochastic variables with sample spaces $\mathcal{S}$ and $\mathcal{T}$, respectively, then

$$H(X,Y) = H(X) + H(Y|X).$$

Proof: By Definition 5.3,

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x,y)) \\
&= -\sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x)p(y|x)) \\
&= -\sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x)) - \sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(y|x)) \\
&= -\sum_{x\in\mathcal{S}} p(x)\log(p(x)) - \sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(y|x)) \\
&= H(X) + H(Y|X)
\end{aligned}$$

■

By use of Theorem 5.6 the following corollaries showing bounds for entropies are established.

Corollary 5.7 (Bound of Joint Entropy)
Let X,Y be discrete stochastic variables. Then

$$H(X,Y) \leq H(X) + H(Y).$$

Proof: This follows directly from the fact that $H(Y) \geq H(Y|X)$.

■

Corollary 5.8 (Entropy of Function of Variable)
Let X be a discrete stochastic variable with sample space $\mathcal{S}$ and f be a deterministic function defined on $\mathcal{S}$. Then

$$H(X) \geq H(f(X)).$$

Proof: Since $H(f(X)|X) = 0$ and

$$H(X, f(X)) = H(X) + H(f(X)|X) = H(f(X)) + H(X|f(X))$$

it follows that

$$H(X, f(X)) = H(X) = H(f(X)) + H(X|f(X)) \quad \Rightarrow \quad H(X) \geq H(f(X))$$

■

Corrollary 5.8 shows that a deterministic function can only maintain or decrease the entropy and therefore the uncertainty of a variable.

5.1.1 Differential Entropy

Up until this point only discrete variables have been discussed. Definitions analogous to Definitions 5.1 and 5.2 for continuous stochastic variables are here introduced.

Definition 5.9 (Differential Entropy and Negentropy)
Let X be a continuous stochastic variable with sample space $\mathcal{S}$ and probability density function p. The differential entropy of X is

$$H_d(X) = -\int_{\mathcal{S}} p(x) \log(p(x)) \, \mathrm{d}x,$$

Definition 5.10 (Differential Negentropy)
Let X be a continuous stochastic variable with sample space $\mathcal{S}$ and probability density function p. The differential negentropy of X is

$$N_d(X) = H_{d,\sigma} - H_d(X),$$

where $H_{d,\sigma}$ is the entropy of a Gaussian stochastic variable with variance equal to the variance σ^2 of X.

Notice that differential negentropy compares the entropy of a stochastic variable to that of the Gaussian distribution which maximises the entropy under a variance constraint.

Definitions 5.3 and 5.5 are equivalent for continuous stochastic variables with sums over sample spaces replaced with integrals over sample spaces. Theorem 5.6 and Corollary 5.7 are likewise equivalent for continuous stochastic variables.

Differential entropy is, in contrast to Shannon entropy, not non-negative and in fact has no lower bound. Consider a Gaussian distribution, where the variance approaches 0. The differential entropy of this diverges toward $-\infty$.

From this point on, all information theoretic quantities are introduced in both the discrete and differential versions. The differential versions are denoted with a subscripted d.

5.2 Kullback-Leibler Divergence

Consider the need to describe similarity of one or more distributions non-parametrically. Definition 5.11 provides a method for doing this.

Definition 5.11 (Kullback-Leibler (KL) Divergence)
Given two probability distributions p, q over a sample space $\mathcal{S}$, the discrete Kullback-Leibler divergence from q to p is

$$D(p||q) = \sum_{x \in \mathcal{S}} p(x) \log \left(\frac{p(x)}{q(x)} \right),$$

while the differential Kullback-Leibler divergence from q to p is

$$D_d(p||q) = \int_{\mathcal{S}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x.$$

In Definition 5.11 the conventions $0 \log\left(\frac{0}{a}\right) \triangleq 0$, $0 \log\left(\frac{0}{0}\right) \triangleq 0$ and $a \log\left(\frac{a}{0}\right) \triangleq \infty$ are used. It should be noted that the KL divergence is not a distance, as it is obviously not symmetric and it does not satisfy the triangle inequality. It can however be interpreted as the non-similarity between two distributions.

Observe that the negentropy in Definition 5.2 is actually the KL divergence from the discrete uniform distribution q to a discrete distribution p. Definition 5.10 is similarly the KL divergence from a Gaussian distribution with variance σ^2 to a continuous distribution p with variance σ^2.

A drawback of using the KL divergence as a measure of non-similarity is that

$$\exists x | p(x) \neq 0, q(x) = 0 \quad \Rightarrow \quad D(p||q) = \infty.$$

5.3 Mutual Information

Mutual information is a special case of KL divergence which quantifies dependency between two stochastic variables.

Definition 5.12 (Mutual Information (MI))
Let X and Y be stochastic variables with joint probability distribution $p(x, y)$, marginal probability distributions $p(x)$ and $p(y)$ and sample spaces $\mathcal{S}$ and $\mathcal{T}$, respectively. The mutual information between discrete X and Y is

$$I(X;Y) = \sum_{x \in \mathcal{S}, y \in \mathcal{T}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) = D(p(x,y)||p(x)p(y)),$$

and the differential mutual information between continuous X and Y is

$$I_d(X;Y) = \int_{\mathcal{S}} \int_{\mathcal{T}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \mathrm{d}x = D_d(p(x,y)||p(x)p(y)).$$

Notice that the MI is symmetric and can be interpreted as the reduction of uncertainty of X given knowledge of Y and vice versa, as the following corollary shows.

Corollary 5.13
Let X and Y be discrete stochastic variables with sample spaces $\mathcal{S}$ and $\mathcal{T}$, respectively, then

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Proof: By Definition 5.12

$$\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= \sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y) \log\left(\frac{p(x|y)}{p(x)}\right) \\
&= -\sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x)) + \sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x|y)) \\
&= -\sum_{x\in\mathcal{S}} p(x)\log(p(x)) - \left(-\sum_{x\in\mathcal{S},y\in\mathcal{T}} p(x,y)\log(p(x|y))\right) \\
&= H(X) - H(X|Y)
\end{aligned}$$

The second part of the proof follows trivially from symmetry.

■

Corollary 5.14 (Bounds of Mutual Information)
Let X and Y be discrete stochastic variables with sample spaces $\mathcal{S}$ and $\mathcal{T}$, respectively. Then

$$0 \leq I(X;Y) \leq \log(\min(|\mathcal{S}|, |\mathcal{T}|)).$$

Proof: Since $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ it follows that

$$I(X;Y) \leq H(X) \leq \log(|\mathcal{S}|) \quad \wedge \quad I(X;Y) \leq H(Y) \leq \log(|\mathcal{T}|),$$

which implies that

$$I(X;Y) \leq \log(\min(|\mathcal{S}|, |\mathcal{T}|)),$$

with equality achieved if and only if X and Y both are deterministically determined from the other and both variables are uniformly distributed on the same sample space.

The lower bound is reached when X and Y are independent such that $H(X|Y) = H(X)$ and $H(Y|X) = H(Y)$.

■

Notice that the differential MI can be rewritten such that $I_d(X;Y) = H_d(X) + H_d(Y) - H_d(X,Y)$, which shows that differential MI is not upper bounded – if X and Y can be deterministically determined from one another, the joint entropy will be $-\infty$ resulting in $I_d(X;Y) = \infty$. This in turn means that the differential self-information $I_d(X;X) = \infty$ in contrast to the discrete where $I(X;X) = H(X)$.

A conditional version of mutual information can be defined from conditional entropy.

Definition 5.15 (Conditional Mutual Information (CMI))
Let X, Y and Z be stochastic variables. The conditional mutual information (CMI) between discrete X, Y and Z is

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

The conditional mutual information between continuous X, Y and Z is analogously defined with H replaced with H_d.

5.3.1 Transfer Entropy

This section is based upon Schreiber [2000] and Baboukani et al. [2020]. When treating time varying variables it may become relevant to describe how much information the past of one of these variables contains about the future of another variable. To this end transfer entropy which relies upon CMI is introduced.

Definition 5.16 (Transfer Entropy (TE))
Let $X = (X_1, X_2, \ldots, X_N)$ and $Y = (Y_1, Y_2, \ldots, Y_N)$ be stationary stochastic processes and denoted source variable and target variable, respectively. Then the Transfer entropy from Y to X is

$$T_{Y\to X}(n) = I(X_n; Y_1, \ldots, Y_{n-1}|X_1, \ldots, X_{n-1}).$$

TE can rewritten into an expression of entropies:

$$T_{Y\to X}(n) = H(X_n|X_1, \ldots, X_{n-1}) - H(X_n|Y_1, \ldots, Y_{n-1}, X_1, \ldots, X_{n-1}).$$

The transfer entropy from the variable Y to X is thus the reduction in uncertainty of X that the inclusion of the past of Y can provide in addition to the reduction which the past of X provides.

The transfer entropy suffers from the fact that information could stem from multiple other sources than just from Y to X, for example a third process Z through which information from Y to X can be conveyed.

Definition 5.17 (Conditional Transfer Entropy (CTE))
Let $X = (X_1, X_2, \ldots, X_N)$,$Y = (Y_1, Y_2, \ldots, Y_N)$ and $Z = (Z_1, Z_2, \ldots Z_N)$ be stationary stochastic processes. Then the conditional transfer entropy from Y to X conditioned on Z is

$$T_{Y\to X|Z}(n) = I(X_n; Y_1, \ldots, Y_{n-1}|X_1, \ldots, X_{n-1}, Z_1, \ldots, Z_{n-1}).$$

CTE can be rewritten into an expression of entropies:

$$\begin{aligned} T_{Y\to X|Z}(n) = {} & H(X_n|X_1, \ldots, X_{n-1}, Z_1, \ldots, Z_{n-1}) \\ & - H(X_n|Y_1, \ldots, Y_{n-1}, X_1, \ldots, X_{n-1}, Z_1, \ldots, Z_{n-1}). \end{aligned}$$

5.4 Generalisations of Mutual Information

While MI expresses reduction in uncertainty of a variable given information about another and therefore dependency between these two variables it fails to specify a generalisation of this to more than two variables. This section treats the problem of quantifying dependency between more than two variables in an information theoretic framework.

5.4.1 Redundancy and Synergy

Quantification of dependencies between more than two variables conventionally relies upon a classification of information into unique, redundant and synergistic. In Williams and Beer [2010] an analysis of the partial information partitions of three stochastic variables into redundant and synergistic information is performed seeking to characterise partial information combinations for three variables. It is further noted in Williams and Beer [2010] that unique information can be described as a combination of redundancy and synergy. This characterisation will not be expanded in its full length here but is summarised as follows:

- Stochastic variables which partially or entirely provide the same information about the outcome of some stochastic variable are characterised as redundant and the redundancy is the amount of information that they share about that variable.

- Stochastic variables which in conjunction provide information which neither provide individually about the outcome of some stochastic variable are characterised as synergistic and the synergy is the additional amount of information provided about that variable.

In Sections 5.4.2, 5.4.3 and 5.4.4 generalisations of MI which seek to characterise and capture the dependency between more than two discrete stochastic variables are presented. In order to determine how well these generalisations quantify the amount of synergy and/or redundancy in a collection of stochastic variables, an average redundancy and an average synergy in certain subsets of a set of stochastic variables are defined.

Definition 5.18 (Average Redundancy and Synergy)
Consider a set of discrete stochastic variables $X_1, \ldots, X_n$. The average pairwise redundancy of $X_1, \ldots, X_n$ is

$$\overline{I}_r(X_1, \ldots, X_n) = \frac{1}{n^2 - n} \sum_{i=1}^{n} \sum_{j \neq i} I(X_i; X_j),$$

and the average synergy at the n'th level of $X_1, \ldots, X_n$ is

$$\overline{I}_{s,n}(X_1, \ldots, X_n) = \frac{1}{n^2 - n} \sum_{i=1}^{n} \sum_{j \neq i} I(X_i; X_j | \{X_1, \ldots, X_n\} \backslash \{X_j, X_i\}).$$

For continuous stochastic variables I is replaced by I_d.

The average redundancy and synergy can be rewritten into expressions of entropies:

$$\begin{aligned}
\overline{I}_r(X_1, \ldots, X_n) &= \frac{1}{n} \sum_{i=1}^{n} \left(H(X_i) - \frac{1}{n-1} \sum_{j \neq i} H(X_i | X_j) \right) \\
\overline{I}_{s,n}(X_1, \ldots, X_n) &= \frac{1}{n} \Bigg[\frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i} H(X_i | \{X_1, \ldots, X_n\} \backslash \{X_j, X_i\}) \\
&\quad - \sum_{k=1}^{n} H(X_k | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n) \Bigg].
\end{aligned}$$

The quantities in Definition 5.18 are interpreted as follows:

- Average redundancy is the average reduction in uncertainty between distinct pairs in the set when one of the variables in the pair is given. It is thus the average redundancy of all pairs consisting of two different variables from the set.
- Average synergy is the average additional reduction in uncertainty in any variable when given all $n-1$ other variables compared to when given any $n-2$

other variables. It is thus the average additional reduction in uncertainty of a subset containing $n-1$ of the variables when adding the n'th variable.

Two types of sets of stochastic variables are introduced and used as examples of redundant and synergistic relations.

Definition 5.19 (n-bit Copy and XOR)
Let $X_1, \ldots, X_n$ be discrete stochastic variables.

- If X_1 is Bernoulli distributed with parameter $p = 1/2$ and $X_1 = \ldots = X_n$ then the set $X_1, \ldots, X_n$ is said to be an n-bit copy. Any one variable is deterministically determined given any one other variable.
- If $X_1, \ldots, X_{n-1}$ are i.i.d. Bernoulli variables with parameter $p = 1/2$ and $X_n = \sum_{k=1}^{n} X_k \pmod 2$, then the set $X_1, \ldots, X_n$ is said to be an n-bit XOR. Any one variable is only reduced in uncertainty given all other variables and is then deterministically determined.

The n-bit copy and n-bit XOR contain redundant and synergistic dependency, respectively.

Below two sets of Bernoulli variables which depending on a parameter η can express a varying degree of redundancy or synergy, respectively, are presented.

Definition 5.20 (Discrete Partial n-bit Copy and XOR)
Let $X_1, \ldots, X_n$ be a set of Bernoulli variables.

- If the probability mass function p of $X_1, \ldots, X_n$ is given by

$$p(x_1, \ldots, x_n) = \begin{cases} (1 + (n-2)\eta/2)/n & \text{for} \quad x_1 = \ldots = x_n \\ (1-\eta)/n & \text{for} \quad x_i \neq x_j, \forall i,j \in \{1, \ldots, n\} \end{cases}$$

  with $\eta \in [0; 1]$, then $X_1, \ldots, X_n$ is called a partial n-bit copy.
- If $X_1, \ldots, X_{n-1}$ are i.i.d. with parameter $p = 1/2$ and X_n is distributed with parameter p given by

$$p = \frac{1}{2} + \frac{\eta}{2}(-1)^{r+1},$$

  where $\eta \in [0; 1]$ and $r = \sum_{i=1}^{k} X_i \pmod 2$ then $X_1, \ldots, X_n$ is called a partial n-bit XOR.

The stochastic variables in Definition 5.20 allow analysis of variables which gradually change from being independent to either redundant or synergistic.

In Figures 5.1 and 5.2 the average redundancies and synergies from Definition 5.18 can be seen for simulated examples of partial 3-bit copies and XORs from Definition 5.20. In Figure 5.3 the difference between average redundancy and synergy for the same simulations is shown.

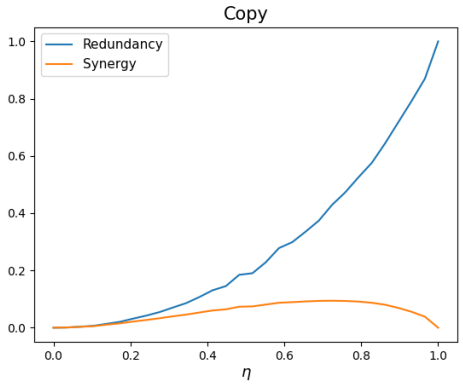


Figure 5.1: Average pairwise redundancy and third level synergy of a partial 3-bit copy for $\eta \in [0; 1]$.

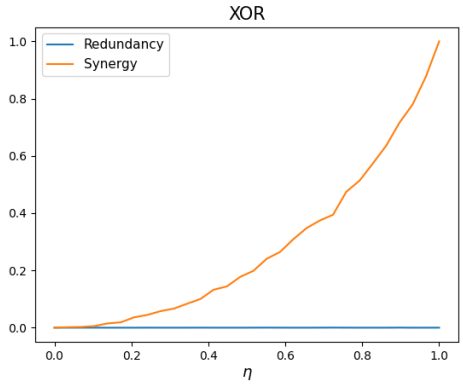


Figure 5.2: Average pairwise redundancy and third level synergy of a partial 3-bit XOR for $\eta \in [0; 1]$.

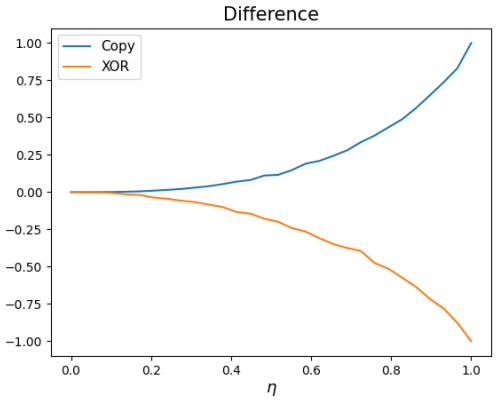


Figure 5.3: Difference between average pairwise redundancy and third level synergy of a partial 3-bit copy and XOR, respectively.

The generalisations of MI to more than two variables presented in the following sections are compared to the average redundancy and synergy to evaluate how well they detect these different types of dependency.

5.4.2 Interaction Information

Interaction information is a generalisation of MI to more than two variables.

Definition 5.21 (Interaction Information)
Let $\mathcal{V} = \{X_1, \ldots, X_n\}$ be a set of stochastic variables which have joint probability distribution $p(x_1, \ldots, x_n)$. The interaction information between discrete $X_1, \ldots, X_n$ is

$$I_{inter}(X_1; \ldots; X_n) = \sum_{\mathcal{T} \subseteq \mathcal{V}} (-1)^{|\mathcal{T}|-1} H(\mathcal{T}).$$

For continuous stochastic variables H is replaced by H_d.

Interaction information can in contrast to MI assume negative values. It is immediately clear that $I_{inter}(X) = H(X)$ and $I_{inter}(X_1; X_2) = I(X_1; X_2)$. For three variables it less clear how to interpret the interaction information.

Definition 5.22
Let X_1, X_2 and X_3 be discrete stochastic variables. If $I_{inter}(X_1; X_2; X_3) > 0$ the variables are said to have a redundant relationship, while if $I_{inter}(X_1; X_2; X_3) < 0$ the variables are said to have a synergistic relationship.

In Example 5.23 is shown interaction information for specific sets of stochastic variables.

Example 5.23 (Three Variable Interaction Information)
Consider a set $\mathcal{V}$ of three discrete stochastic variables X_1, X_2 and X_3. Their interaction information is then given by

$$\begin{aligned} I_{inter}(X_1; X_2; X_3) &= \sum_{\mathcal{T} \subseteq \mathcal{V}} (-1)^{|\mathcal{T}|-1} H(\mathcal{T}) \\ &= H(X_1) + H(X_2) + H(X_3) \\ &\quad - (H(X_1,X_2) + H(X_1,X_3) + H(X_2,X_3)) \\ &\quad + H(X_1,X_2,X_3). \end{aligned}$$

Redundant relationship
Consider a 3-bit copy. Then $H(X_1) = H(X_1,X_2) = H(X_1,X_2,X_3)$ and it follows that

$$\begin{aligned} I_{inter}(X_1; X_2; X_3) &= 1 + 1 + 1 \\ &\quad - (1 + 1 + 1) \\ &\quad + 1 \\ &= 1. \end{aligned}$$

The variables contribute the same information and are therefore redundant.

Synergistic relationship
Consider a 3-bit XOR. Then $H(X_1,X_2,X_3) = 2$ and it follows that

$$\begin{aligned} I_{inter}(X_1; X_2; X_3) &= 1 + 1 + 1 \\ &\quad - (2 + 2 + 2) \\ &\quad + 2 \\ &= -1. \end{aligned}$$

The variables exhibit synergistic relationship since none of them are reduced in uncertainty from knowledge of a single of the other two variables but they are deterministically determined from knowledge of the other two.

While interaction information from Definition 5.22 has a defined interpretation for three variables this does not generalise to more than three variables. It is sometimes interpreted as the information shared by all included variables which is not shared by any subset of these variables Williams and Beer [2010].

5.4.3 Total and Dual Total Correlation

Total correlation (TC) and dual total correlation (DTC) are non-negative generalisations of MI.

Definition 5.24 (Total Correlation (TC))
Let $X_1, \ldots, X_n$ be stochastic variables with joint probability distribution $p(x_1, \ldots, x_n)$ and marginal probability distributions $p(x_1), \ldots, p(x_n)$. The total correlation of discrete $X_1, \ldots, X_n$ is

$$C(X_1, \ldots, X_n) = D\left(p\left(x_1, \ldots, x_n\right) || \prod_{k=1}^{n} p(x_k)\right),$$

where $D(\cdot||\cdot)$ is the KL divergence. The differential total correlation C_d between continuous $X_1, \ldots, X_n$ is defined analogously by switching D with D_d.

TC is as such the difference as quantified by the KL divergence between the joint probability density and the product of the marginals similar to MI. From the properties of the KL divergence it follows that TC is non-negative and is minimised when $X_1, \ldots, X_n$ are independent such that the joint probability function of $X_1, \ldots, X_n$ equals the product of the marginals.

Corollary 5.25
Let $X_1, \ldots, X_n$ be stochastic variables. If $X_1, \ldots, X_n$ are discrete, then

$$0 \leq C(X_1, \ldots, X_n) \leq \sum_{k=1}^{n} H(X_k).$$

If $X_1, \ldots, X_n$ are continuous, then

$$0 \leq C_d(X_1, \ldots, X_n).$$

Proof: Total correlation for discrete variables is rewritten to be expressed as entropies:

$$\begin{aligned}
C(X_1, \ldots, X_n) &= D\left(p(x_1, \ldots, x_n) || \prod_{k=1}^{n} p(x_k)\right) \\
&= \sum_{x_1 \in \mathcal{S}_1, \ldots, x_n \in \mathcal{S}_n} p(x_1, \ldots, x_n) \log\left(\frac{p(x_1, \ldots, x_n)}{\prod_{k=1}^{n} p(x_k)}\right) \\
&= \sum_{x_1 \in \mathcal{S}_1, \ldots, x_n \in \mathcal{S}_n} p(x_1, \ldots, x_n) \left(\log\left(p(x_1, \ldots, x_n)\right) - \log\left(\prod_{k=1}^{n} p(x_k)\right)\right) \\
&= -H(X_1, \ldots, X_n) - \sum_{x_1 \in \mathcal{S}_1, \ldots, x_n \in \mathcal{S}_n} p(x_1, \ldots, x_n) \sum_{k=1}^{n} \log(p(x_k)) \\
&= \sum_{k=1}^{n} H(X_k) - H(X_1, \ldots, X_n). \qquad (5.1)
\end{aligned}$$

Both bounds are established through the fact that $H(X_1, \ldots, X_n) \leq \sum_{k=1}^{n} H(X_k)$, $H(X_1, \ldots, X_n) \geq 0$ and $H(X_k) \geq 0$ for $k \in \{1, \ldots, n\}$. If $X_1, \ldots, X_n$ are continuous, then Equation (5.1) is still valid with differential entropy. From the fact that $\sum_{k=1}^{n} H_d(X_k) \geq H_d(X_1, \ldots, X_n)$ it from the continuous version of Corollary 5.7 it follows that $C_d(X_1, \ldots, X_n) \geq 0$. The lack of upper bound is seen if $H_d(X_k)$ is finite for $k \in \{1, \ldots, n\}$ and $H_d(X_1, \ldots, X_n) = -\infty$. Then $C_d(X_1, \ldots, X_n) = \infty$.

■

Definition 5.26 (Dual Total Correlation (DTC))
Let $X_1, \ldots, X_n$ be stochastic variables. The dual total correlation of discrete $X_1, \ldots, X_n$ is

$$B(X_1, \ldots, X_n) = H(X_1, \ldots, X_n) - \sum_{k=1}^{n} R_k(X_1, \ldots, X_n),$$

where

$$R_k(X_1, \ldots, X_n) = H(X_k | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n).$$

The differential dual total correlation of continuous $X_1, \ldots, X_n$ is defined analogously by switching H with H_d.

The term $R_k(X_1, \ldots, X_n)$ in Definition 5.26 is referred to as the residual entropy of X_k and quantifies how much information can only be obtained from observing X_k. The differential version of R_k is denoted $R_{k,d}$.

Corollary 5.27
Let $X_1, \ldots, X_n$ be stochastic variables. If $X_1, \ldots, X_n$ are discrete, then

$$0 \leq B(X_1, \ldots, X_n) \leq H(X_1, \ldots, X_n).$$

If $X_1, \ldots, X_n$ are continuous, then

$$0 \leq B_d(X_1, \ldots, X_n).$$

Proof: Since joint entropy is related to conditional entropy through the chain rule in Theorem 5.6:

$$H(X_1, \ldots, X_n) = \sum_{k=1}^{n} H(X_k | X_{k-1}, \ldots, X_1),$$

it follows that

$$H(X_1, \ldots, X_n) \geq \sum_{k=1}^{n} H(X_k | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n) = R_k, \tag{5.2}$$

which together with the fact that $H(X_k | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n) \geq 0$ for $k \in \{1, \ldots, n\}$ proves both the upper and lower bound for discrete variables. If $X_1, \ldots, X_n$ are continuous, Equation (5.2) is still valid with differential entropy. From Equation (5.2) it follows that $B_d(X_1, \ldots, X_n) \geq 0$. The lack of upper bound is seen if $H_d(X_1, \ldots, X_n)$ is finite and $\exists k \in \{1, \ldots, n\}(R_{k,d}(X_1, \ldots, X_n) = -\infty)$. Then $B_d(X_1, \ldots, X_n) = \infty$.

■

5.4.4 O-information

In Rosas et al. [2019] a generalisation of MI to more than two variables is presented, namely information about organisational structure (O-information). It relies upon the TC and DTC.

Definition 5.28 (O-information)
Let $X_1, \ldots, X_n$ be stochastic variables. The O-information of discrete $X_1, \ldots, X_n$ is defined as

$$\mathcal{O}(X_1, \ldots, X_n) = C(X_1, \ldots, X_n) - B(X_1, \ldots, X_n).$$

The differential O-information of continuous $X_1, \ldots, X_n$ is defined analogously by switching B and C with B_d and C_d, respectively.

The O-information is the difference between the TC and DTC and can be rewritten into a simpler sum of entropies:

$$\begin{aligned}\mathcal{O} &= C(X_1, \ldots, X_n) - B(X_1, \ldots, X_n) \\ &= (n-2)H(X_1, \ldots, X_n) + \sum_{j=1}^{n} \left(H(X_j) - H(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)\right).\end{aligned}$$

Two interesting properties of the O-information are that $\mathcal{O}(X_1, X_2) = 0$ and $\mathcal{O}(X_1, X_2, X_3) = I_{inter}(X_1; X_2; X_3)$. This means that the synergy and redundancy between only two variables are equal and the interaction information coincides with the O-information for three variables.

Corollary 5.29
Let X_1, X_2 and X_3 be discrete stochastic variables. Then

$$\begin{aligned}\mathcal{O}(X_1) &= 0, \\ \mathcal{O}(X_1, X_2) &= 0, \\ \mathcal{O}(X_1, X_2, X_3) &= I_{inter}(X_1; X_2; X_3).\end{aligned}$$

Proof:

O-information of one variable:

$$\begin{aligned}\mathcal{O}(X_1) &= C(X_1) - B(X_1)\\ &= H(X_1) - H(X_1) - (H(X_1) - H(X_1))\\ &= 0\end{aligned}$$

O-information of two variables:

$$\begin{aligned}\mathcal{O}(X_1, X_2) &= C(X_1, X_2) - D(X_1, X_2)\\ &= H(X_1) + H(X_2) - H(X_1, X_2)\\ &\quad - (H(X_1, X_2) - H(X_1|X_2) - H(X_2|X_1))\\ &= 2H(X_1, X_2) - 2H(X_1, X_2)\\ &= 0.\end{aligned}$$

O-information of three variables:

$$\begin{aligned}\mathcal{O}(X_1, X_2, X_3) &= C(X_1, X_2, X_3) - D(X_1, X_2, X_3)\\ &= H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) - (H(X_1, X_2, X_3)\\ &\quad - H(X_1|X_2, X_3) - H(X_2|X_1, X_3) - H(X_3|X_1, X_2))\\ &= H(X_1) + H(X_2) + H(X_3)\\ &\quad - H(X_1, X_2)\\ &\quad - (H(X_1, X_2, X_3) - H(X_1|X_2, X_3) - H(X_2|X_1, X_3))\\ &= H(X_1) + H(X_2) + H(X_3)\\ &\quad - (H(X_1, X_2) + H(X_2, X_3))\\ &\quad + H(X_2|X_1, X_3)\\ &= H(X_1) + H(X_2) + H(X_3)\\ &\quad - (H(X_1, X_2) + H(X_1, X_3)) + H(X_2, X_3)\\ &\quad + H(X_1, X_2, X_3)\\ &= I_{inter}(X_1; X_2; X_3).\end{aligned}$$

■

The proof of Corollary 5.29 is analogous for continuous stochastic variables.

The interpretation of the O-information is established through the following definition.

Definition 5.30
If $\mathcal{O}(X_1, \ldots, X_n) > 0$ then $X_1, \ldots, X_n$ are said to be redundancy dominated, while if $\mathcal{O}(X_1, \ldots, X_n) < 0$ then $X_1, \ldots, X_n$ are said to be synergy dominated.

Definition 5.30 gives an interpretation of the relationship between a set of stochastic variables based upon the O-information which generalises to more than

three variables in contrast to Definition 5.22. Notice however that Definition 5.30 only concerns whether a system is more redundant or synergistic – if both TC and DTC are positive but equal it results in $\mathcal{O} = 0$, which is not to be interpreted as a sign of absence of dependency.

5.4.5 Examples

In this section previously presented methods of quantifying dependency between more than two variables are tested on 2-, 3- and 4-bit copies and XORs as defined in Definition 5.19. The methods include

- average redundancy ($\overline{I}_r$),
- average synergy ($\overline{I}_{s,n}$),
- interaction information (I_{inter}),
- total correlation (C),
- dual total correlation (B), and
- O-information ($\mathcal{O}$).

The results from applying the above listed methods for quantifying multivariate dependency on different sets of stochastic variables can be seen in Table 5.1. The first column for example shows the results of applying the average redundancy in Definition 5.18 on 2-, 3- and 4-bit copies and XORs. The table, although based upon specific examples, allows a superficial comparison between the presented methods.

	$\overline{I}_r$	$\overline{I}_{s,n}$	I_{inter}	C	B	$\mathcal{O}$
2-bit copy	1	1	1	1	1	0
2-bit XOR	1	1	1	1	1	0
3-bit copy	1	0	1	2	1	1
3-bit XOR	0	1	-1	1	2	-1
4-bit copy	1	0	1	3	1	2
4-bit XOR	0	1	1	1	3	-2

Table 5.1: Multivariate methods for quantifying dependency applied to 2-, 3-, and 4-bit copies and XORs.

The following observations can be made from Table 5.1:

- None of the methods can distinguish between a 2-bit copy and 2-bit XOR since redundancy and synergy for two variables are equivalent.
- As a consequence of Definition 5.18, the average pairwise redundancy and average synergy at the n'th level can not distinguish between n-bit copies or XORs of varying n.
- Interaction information can not not distinguish between a 4-bit copy and XOR.

- TC and DTC successfully distinguish redundancy and synergy for more than two variables in the case of n-bit copies and XORs, although to a different degree than the average quantities.
- O-information successfully determines the dominant type of dependency present in the tested sets of variables.

Notice furthermore that O-information as earlier mentioned is only meant to specify whether a system is dominantly synergistic or redundant. A set of six stochastic variables of which three constitute a 3-bit copy while the remaining three constitute a 3-bit XOR independent of the 3-bit copy has an O-information of 0.

The TC and DTC of simulated partial 3-bit copies and XORs from Definition 5.20, for $\eta \in [0; 1]$ respectively can be seen in Figures 5.4 and 5.5 while Figure 5.6 shows the resulting O-information. Plots are similar to those in Figures 5.1, 5.2 and 5.3, which shows that TC, DTC and O-information succeed in capturing redundancy and synergy in agreement with the quantities in Definition 5.18.

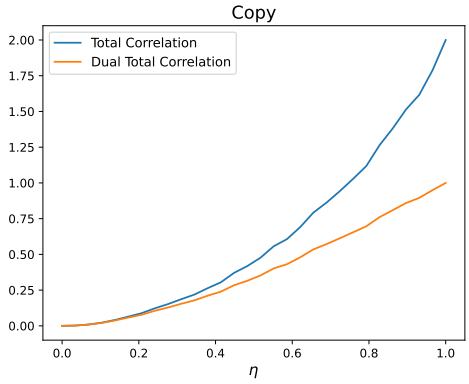


Figure 5.4: TC and DTC of partial 3-bit copy for $\eta \in [0; 1]$.

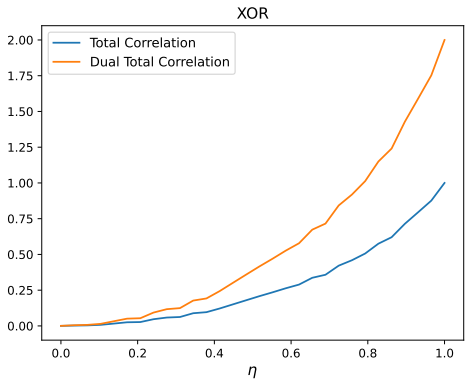


Figure 5.5: TC and DTC of partial 3-bit XOR for $\eta \in [0; 1]$.

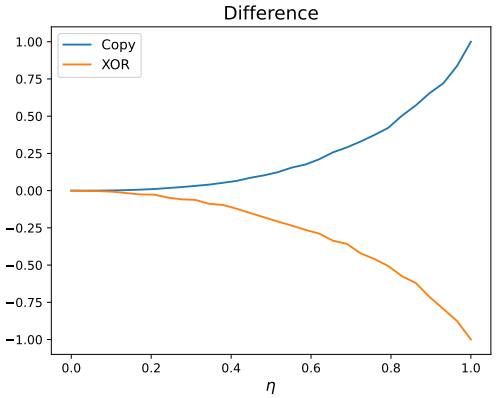


Figure 5.6: O-information of partial 3-bit copy and XOR for $\eta \in [0; 1]$.

6 | Estimation of Entropy and Total Correlation

This chapter presents methods for estimating differential entropy and differential total correlation. The chapter is based upon Gao et al. [2015].

6.1 Differential Entropy

Entropy, whether discrete or continuous, can be estimated through binning. For discrete variables, the bins can be determined from the nature of the data, while it is a free parameter for continuous distributions, which decides upon the step size of a quantisation of the continuous variable. In order to avoid the problem of determining a binning method for continuous variables and the artifacts of using a binning method, a non-parametric estimator of entropy for continuous variables can be used.

Definition 6.1 (*k*-Nearest Neighbour (*k*NN) Entropy Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The kNN entropy estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{H}'_{k\text{NN},k}(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\hat{p}_k\left(\mathbf{x}^{(i)}\right)\right)$$

where

$$\hat{p}_k\left(\mathbf{x}^{(i)}\right) = \frac{k}{N-1}\frac{\Gamma(d/2+1)}{\pi^{d/2}} r_k\left(\mathbf{x}^{(i)}\right)^{-d},$$

and $r_k\left(\mathbf{x}^{(i)}\right)$ is the Euclidean distance to the k'th nearest neighbour of $\mathbf{x}^{(i)}$.

The estimator in Definition 6.1 seeks to estimate the entropy of multiple stochastic variables directly by using a k-nearest neighbor estimate of density.

Definition 6.2 (Unbiased *k*NN Entropy Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The unbiased kNN entropy estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{H}_{k\text{NN},k}(\mathbf{x}) = \hat{H}'_{kNN,k}(\mathbf{x}) - \psi(k) - \log(k),$$

where $\psi(k) = \frac{\mathrm{d}}{\mathrm{d}k}\ln(\Gamma(k))$ is the digamma function and

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}z,$$

is the gamma function.

The correction terms in Definition 6.2 makes the estimate $\hat{H}_{k\text{NN},k}$ asymptotically unbiased.

Theorem 6.3 (Asymptotical Unbiasedness of $\hat{H}_{k\text{NN},k}$)
If $X_1, \ldots, X_d$ are absolutely continuous stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ are N i.i.d. samples of $X_1, \ldots, X_d$, then

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{H}_{k\text{NN},k}(\mathbf{x})\right] = H(X_1, \ldots, X_d).$$

The proof of Theorem 6.3 has been omitted.

In Kraskov et al. [2004] an estimator inspired by the kNN estimator referred to as the KSG estimator of entropy is introduced. This estimator was in Khan et al. [2007] shown to produce good result across various noisy signals of short length and with varying types of dependencies.

Definition 6.4 (KSG Entropy Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The KSG entropy estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{H}_{\text{KSG},k}(\mathbf{x}) = \psi(N) - \psi(k) + \frac{d}{N}\sum_{i=1}^{N}\log(\epsilon_k(\mathbf{x}_i)),$$

where $\epsilon_k(\mathbf{x}_i)$ is twice the max-norm distance to the k'th nearest neighbour of $\mathbf{x}^{(i)}$.

6.2 Differential Total Correlation

Since total correlation can be decomposed into an expression of entropies as seen in the proof of Corollary 5.25, the estimators in Section 6.1 can be used for estimation of total correlation.

Definition 6.5 (*k*NN Total Correlation Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The kNN total correlation estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{C}'_{k\text{NN},k}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\hat{p}_k\left(\mathbf{x}^{(i)}\right)}{\hat{p}_k\left(\mathbf{x}_1^{(i)}\right)\cdots\hat{p}_k\left(\mathbf{x}_d^{(i)}\right)}\right),$$

where

$$\hat{p}_k\left(\mathbf{x}^{(i)}\right) = \frac{k}{N-1}\frac{\Gamma(d/2+1)}{\pi^{d/2}} r_k\left(\mathbf{x}^{(i)}\right)^{-d},$$

and $r_k\left(\mathbf{x}^{(i)}\right)$ is the Euclidean distance to the k'th nearest neighbour of $\mathbf{x}^{(i)}$.

The estimator in Definition 6.5 is like the kNN entropy estimator biased and can be corrected by an offset.

Definition 6.6 (Unbiased *k*NN Total Correlation Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The unbiased kNN total correlation estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{C}_{k\text{NN},k}(\mathbf{x}) = \hat{C}'_{k\text{NN},k}(\mathbf{x}) - (d-1)(\psi(k) - \log(k)).$$

The kNN estimate of total correlation in Definition 6.6 is also asymptotically unbiased.

Theorem 6.7 (Asymptotical Unbiasedness of $\hat{C}_{k\text{NN},k}$)
If $X_1, \ldots, X_d$ are absolutely continuous stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ are i.i.d. samples of X, then

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{C}_{k\text{NN},k}(\mathbf{x})\right] = C(X_1, \ldots, X_d).$$

The proof of Theorem 6.7 has been omitted.

The KSG entropy estimator can furthermore be used for an estimate of total correlation.

Definition 6.8 (KSG Total Correlation Estimator)
Let $X_1, \ldots, X_d$ be stochastic variables and $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ be N i.i.d. samples of $X_1, \ldots, X_d$ such that $\mathbf{x}_j^{(i)}$ is the i'th sample of X_j. The KSG total correlation estimator of $X_1, \ldots, X_d$ is defined as

$$\hat{C}_{\text{KSG},k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - \frac{d-1}{k} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi(n_{\mathbf{x}_j}(i)),$$

where $n_{\mathbf{x}_j}(i)$ is the number of points in $\mathbf{x}_j$ at a distance from $\mathbf{x}_j^{(i)}$ less than or equal to the max-norm distance from $\mathbf{x}_j^{(i)}$ to the k'th-nearest neighbour of $\mathbf{x}^{(i)}$ in $\mathbf{x}$.

The estimators presented above work when the dependency between variables is weak or the number of samples is very large [Gao et al., 2015]. The following theorem shows that the performance of the estimators are influenced by the dimension d, the number of samples N and the magnitude of the true total correlation $C(X_1, \ldots, X_d)$.

Theorem 6.9 (Convergence Rate of Bias of *k*NN $C(\mathbf{x})$ Estimator)
Let $p(\mathbf{x})$ be any absolutely continuous d-dimensional probability distribution, $k \geq 1$ and $\epsilon > 0$. Then in order for $|\hat{C}_{k\text{NN},k}(\mathbf{x}) - C(X_1, \ldots, X_d)| \leq \epsilon$, where $\mathbf{x} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ are N i.i.d. samples of $X_1, \ldots, X_d$, the inequality $N \geq C \exp\left(\frac{C(\mathbf{x})-\epsilon}{d-1}\right) + 1$, where C is a constant which scales with $O\left(\frac{1}{d}\right)$, must be satisfied.

Proof: Consider the following rearrangement:

$$\begin{aligned}
\hat{C}_{k\mathrm{NN},k} &= \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\hat{p}(\mathbf{x}^{(i)})}{\hat{p}_k(\mathbf{x}_1^{(i)})\hat{p}_k(\mathbf{x}_2^{(i)})\cdots\hat{p}_k(\mathbf{x}_k^{(i)})}\right) - (d-1)\gamma_k \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\frac{k}{N-1}\frac{\Gamma(d/2+1)}{\pi^{d/2}}r_k(\mathbf{x}^{(i)})^{-d}}{\prod_{j=1}^{d}\frac{k}{N-1}\frac{\Gamma(1/2+1)}{\pi^{1/2}}r_k(\mathbf{x}_j^{(i)})^{-1}}\right) - (d-1)\gamma_k \\
&= \frac{1}{N}\sum_{i=1}^{N}\Bigg\{\log\left(\frac{k}{N-1}\right) + \log\left(\frac{\Gamma(d/2+1)}{\pi^{d/2}}\right) + \log\Big(r_k(\mathbf{x}^{(i)})^{-d}\Big) \\
&\quad - \sum_{j=1}^{d}\Bigg[\log\left(\frac{k}{N-1}\right) + \log\left(\frac{\Gamma(1/2+1)}{\pi^{1/2}}\right) \\
&\quad + \log\left(\frac{\Gamma(1/2+1)}{\pi^{1/2}}\right) - \log\Big(r_k(\mathbf{x}_j^{(i)})\Big)\Bigg]\Bigg\} - (d-1)\gamma_k \\
&= \frac{1}{N}\Bigg[N\log\left(\frac{k}{N-1}\right) + N\log\left(\frac{\Gamma(d/2+1)}{\pi^{d/2}}\right) - \sum_{i=1}^{N} d\log\Big(r_k(\mathbf{x}^{(i)})\Big) \\
&\quad - Nd\log\left(\frac{k}{N-1}\right) - Nd\log\left(\frac{\Gamma(1/2+1)}{\pi^{1/2}}\right) + \sum_{m=1}^{N}\sum_{j=1}^{d} r_k(\mathbf{x}_j^{(m)})\Bigg] \\
&\quad - (d-1)\gamma_k \\
&= \log\left(\frac{k}{N-1}\right) + \log\left(\frac{\Gamma(d/2+1)}{\pi^{d/2}}\right) - \frac{d}{N}\sum_{i=1}^{N}\log\Big(r_k(\mathbf{x}^{(i)})\Big) \\
&\quad - d\log\left(\frac{k}{N-1}\right) - d\log\left(\frac{\Gamma(1/2+1)}{\pi^{1/2}}\right) + \frac{1}{N}\sum_{m=1}^{N}\sum_{j=1}^{d}\log\Big(r_k(\mathbf{x}_j^{(m)})\Big) \\
&\quad - (d-1)\gamma_k \\
&= (d-1)\log\left(\frac{N-1}{k}\right) + \log\left(\frac{\Gamma(d/2+1)}{(\Gamma(1/2+1))^d}\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\Big(r_k(\mathbf{x}_j^{(i)})\Big) - \frac{d}{N}\sum_{i=1}^{N}\log\Big(r_k(\mathbf{x}^{(i)})\Big) - (d-1)\gamma_k. \qquad (6.1)
\end{aligned}$$

Using the fact that

$$r_k(\mathbf{x}^{(i)}) \geq r_k(\mathbf{x}_j^{(i)}),$$

the following inequality can be established:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\Big(r_k(\mathbf{x}_j^{(i)})\Big) - \frac{d}{N}\sum_{i=1}^{N}\log\Big(r_k(\mathbf{x}^{(i)})\Big) \leq 0, \qquad (6.2)$$

where equality is achieved when $r_k(\mathbf{x}^{(i)}) = r_k(\mathbf{x}_j^{(i)})$ for all $j \in \{1,\ldots,d\}$. Utilising

Equation (6.2) with Equation (6.1):

$$\begin{aligned}
&(d-1)\log\Big(\frac{N-1}{k}\Big) + \log\Bigg(\frac{\Gamma(d/2+1)}{(\Gamma(1/2+1))^d}\Bigg) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\Big(r_k(\mathbf{x}_j^{(i)})\Big) - \frac{d}{N}\sum_{i=1}^{N}\log\Big(r_k(\mathbf{x}^{(i)})\Big) - (d-1)\gamma_k \\
\leq\ &(d-1)\log\Big(\frac{N-1}{k}\Big) + \log\Bigg(\frac{\Gamma(d/2+1)}{(\Gamma(1/2+1))^d}\Bigg) - (d-1)\gamma_k
\end{aligned}$$

Observing that $\psi(k) - \log(k)$ is a monotonously decreasing function Gao et al. [2015] and setting $k = 1$ results in

$$\begin{aligned}
&(d-1)\log\Big(\frac{N-1}{k}\Big) + \log\Bigg(\frac{\Gamma(d/2+1)}{(\Gamma(1/2)+1)^d}\Bigg) - (d-1)(\psi(k) - \log(k)) \\
\leq\ &(d-1)\log(N-1) + \log\Bigg(\frac{\Gamma(d/2+1)}{(\Gamma(1/2)+1)^d}\Bigg) - (d-1)(\psi(1) - \log(1)). \quad (6.3)
\end{aligned}$$

Secondly, using a bound of Γ found in Gao et al. [2015]:

$$\begin{aligned}
\log\Bigg(\frac{\Gamma(d/2+1)}{\Gamma(1/2+1)^d}\Bigg) &= \log(\Gamma(d/2+1)) - d\log(\Gamma(1/2+1)) \\
&< \log\Bigg(\sqrt{2\pi}\Big(\frac{d/2+1/2}{e}\Big)^{d/2+1/2}\Bigg) - d\log\Big(\pi^{1/2}+1\Big) \quad \text{(for large } d\text{)} \\
&= O(d\log(d)).
\end{aligned}$$

Using the above inequality in Equation (6.3) yields

$$\begin{aligned}
\hat{C}_{kNN,k}(\mathbf{x}) &\leq (d-1)\log\Big(\frac{N-1}{k}\Big) + O(d\log(d)) \\
&\leq (d-1)\log(N-1) + O(d\log(d)).
\end{aligned}$$

Enforcing that $\Big|\hat{C}_{kNN,k}(\mathbf{x}) - C(\mathbf{x})\Big| \leq \epsilon$, then

$$N \geq C\exp\left(\frac{C(\mathbf{x}) - \epsilon}{d-1}\right) + 1,$$

where C is a constant that scales like $O(\frac{1}{d})$ [Gao et al., 2015]. ■

Theorem 6.9 shows that the rate of convergence of the kNN total correlation estimator toward the true total correlation is dependent on the number of samples N, the dimension d and the magnitude of $C(X_1, \ldots, X_d)$. The number of samples needed for a good estimate might therefore grow large. A similar result for $I_{KSG,k}(\mathbf{x})$ is found in Gao et al. [2015].

7 | Omega Complexity

Omega complexity (OC) was introduced in Wackermann [1996] as a method of quantifying dependency between EEG signals and it has since been used extensively for this purpose [Zhao et al., 2022; Gaál et al., 2010; Irisawa et al., 2006; Starn et al., 2000; Toth et al., 2009]. In this chapter some properties of the OC are examined and related to the problem of quantifying dependency between EEG signals.

Before defining the OC, the terms dependency function and dependency matrix are introduced, as the OC relies upon these.

Definition 7.1 (Dependency Function)
Let X and Y be stochastic variable. A dependency function is a function $\rho(X, Y) \in \mathbb{R}$, which quantifies how dependent X and Y are on each other. It must satisfy that higher dependency between X and Y results in a higher absolute value of ρ and furthermore that

- $\rho(X, Y) = \rho(Y, X)$,
- $|\rho| \in [0; 1]$,
- X and Y are independent $\Rightarrow \rho = 0$, and
- X and Y are deterministically determined from each other $\Rightarrow |\rho| = 1$.

If more than two variables are treated, their pairwise dependencies can be organised in a matrix.

Definition 7.2 (Dependency Matrix)
Given a set $S = \{S_k\}_{k=1}^n$ of stochastic variables and a dependency function ρ, the matrix $C^{\rho,S} \in \mathbb{R}^{n\times n}$ whose entries are described by

$$c_{ij}^{\rho,S} = \rho(S_i, S_j) \quad \text{for} \quad i, j \in \{1, \dots, n\},$$

is called the dependency matrix of S with respect to ρ. The set

$$\mathcal{C}_{\rho,S}^n = \{C^{\rho,S} \in \mathbb{R}^{n\times n} \,|\, c_{ij}^{\rho,S} = \rho(S_i, S_j), i, j \in \{1, \dots, n\}\},$$

where ρ is any dependency function and $S = \{S_k\}_{k=1}^n$ is any set of n stochastic variables, is referred to as the set of dependency matrices of dimension n.

The dependency matrices are all real and symmetric but not necessarily positive semi-definite and it is difficult to describe properties of them, since the characteristics of S and ρ can vary greatly. Notice, that $\mathcal{C}^n_{\rho,S}$ contains all correlation matrices of dimension n.

In Wackermann [1996] the OC was introduced. This project will present a version of OC modified from the one presented in Wackermann [1996] such that it has range $[0; 1]$ and a high OC is interpreted as high dependency and vice versa – this version is also found in Baboukani et al. [2018]. The results found in this chapter are therefore specific to the version of the OC presented below, however preliminary tests indicate that some of the results may be true for the OC in Wackermann [1996] with small modifications.

Definition 7.3 (Omega Complexity)
Let $S = \{S_k\}_{k=1}^n$ be a set of stochastic variables, ρ a dependency function and $C^{\rho,S}$ the dependency matrix. The omega complexity of S with respect to ρ is defined as

$$\Omega(\rho; S) = 1 + \frac{\sum_{k=1}^n \overline{\lambda}_k \log\left(\overline{\lambda}_k\right)}{\log(n)}, \quad \text{where} \quad \overline{\lambda_k} = \frac{|\lambda_k|}{\sum_{m=1}^n |\lambda_m|},$$

and $\{\lambda_k\}_{k=1}^n$ are the eigenvalues of $C^{\rho,S}$.

Notice, that the eigenvalues of $C^{\rho,S}$ are real, since $C^{\rho,S}$ is Hermitian. The choice of ρ furthermore influences whether $C^{\rho,S}$ is positive semi-definite.

7.1 Analysis and Examples

This section seeks to uncover some of the properties of the OC through analysis and examples.

Theorem 7.4 (Properties of Omega Complexity)
The omega complexity

a) has range $[0; 1]$,

b) is maximised when $\text{rank}(C^{\rho,S}) = 1$ such that $\{\overline{\lambda}_k\}_{k=1}^n = \{1, 0, \ldots, 0\}$,

c) is minimised when $C^{\rho,S}$ is the identity matrix such that $\{\overline{\lambda}_k\}_{k=1}^n = \{1/n, \ldots, 1/n\}$, and

d) is invariant to scaling of S if ρ is invariant to scaling of S.

Proof:

a) Observe that since the absolute value of the n eigenvalues are ℓ_1-normalised, $\sum_{k=1}^{n} \overline{\lambda}_k \log\left(\overline{\lambda}_k\right)$ is analogous to the negative entropy of these. The entropy of a discrete stochastic variable with n possible events has range $[0; \log(n)]$ and this is normalised by $\log(n)$.

b) If $\text{rank}(C^{\rho,S}) = 1$ then $\text{null}(C^{\rho,S}) = n - 1$ which results in an eigenvalue of 0 with multiplicity $n - 1$ and therefore only one non-zero eigenvalue. This eigenvalue is then normalised through $\overline{\lambda}_1 = |\lambda_1| / \sum_{m=1}^{n} |\lambda_m| = 1$.

c) Since a uniform distribution of the eigenvalues minimises the OC, the minimum is achieved with $\{|\overline{\lambda}_k|\}_{k=1}^{n} = \{1/n, \ldots, 1/n\}$. As $C^{\rho,S}$ is a real, symmetric matrix the spectral theorem states that $C^{\rho,S}$ is diagonalisable. Hence a decomposition exists such that

$$C^{\rho,S} = PDP^{-1},$$

where D is a diagonal matrix with the eigenvalues of $C^{\rho,S}$ on the diagonal. Assume that $C^{\rho,S}$ has one eigenvalue of 1 with multiplicity n such that the decomposition can be rewritten as

$$C^{\rho,S} = PIP^{-1},$$

which implies that

$$C^{\rho,S}C^{\rho,S} = PIP^{-1}PIP^{-1} = I.$$

That is, $C^{\rho,S}$ is an involutory matrix. Since all symmetric involutory matrices are orthogonal and the diagonal of $C^{\rho,S}$ consists only of ones it follows that $C^{\rho,S} = I$.

d) If ρ is scale invariant such that

$$\rho(as_i, as_j) = \rho(s_i, s_j)$$

for any $a \in \mathbb{R}\backslash 0$, then the OC is trivially scale invariant.

■

Notice that while properties b) and c) concern the maximisation and minimisation of the OC, respectively, only c) specifies a unique correlation matrix which achieves this while the only requirement for maximisation is that all columns in the correlation matrix should be linearly dependent. This suggests that the OC might confound the dependency of two different sets of signals as the same.

The OC varies as a function of the covariance between two Bernoulli variables, as is shown in Example 7.5.

Example 7.5 (Covarying Bernoulli Variables)
Consider two stochastic variables $X = X_1, X_2$ with covariance ξ which can be normalised to a Pearson correlation coefficient $\rho_p \in [-1; 1]$. The dependency matrix $C^{\rho_p, X}$ is then given as

$$C^{\rho_p, X} = \begin{bmatrix} 1 & \rho_p \\ \rho_p & 1 \end{bmatrix}.$$

The eigenvalues of $C^{\rho, X}$ are then given as the roots of the characteristic polynomial of $C^{\rho, X}$:

$$\det\left(\begin{bmatrix} 1-\lambda & \rho_p \\ \rho_p & 1-\lambda \end{bmatrix}\right) = 0 \qquad \Longrightarrow \qquad \lambda = 1 \pm \rho_p$$

The OC is therefore

$$\Omega(\rho_p; X) = 1 + \left(\frac{1+\rho_p}{2}\right)\log\left(\frac{1+\rho_p}{2}\right) + \left(\frac{1-\rho_p}{2}\right)\log\left(\frac{1-\rho_p}{2}\right), \tag{7.1}$$

where $0\log(0) \triangleq 0$. Differentiating the OC with respect to ρ_p:

$$\frac{\mathrm{d}}{\mathrm{d}\rho_p}\Omega(\rho_p; X) = \frac{1}{2}\log\left(\frac{1+\rho_p}{1-\rho_p}\right) = \begin{cases} < 0, & \text{for } \rho_p < 0 \\ 0, & \text{for } \rho_p = 0 \\ > 0, & \text{for } \rho_p > 0 \end{cases} \tag{7.2}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\rho_p^2}\Omega(\rho_p; X) = \frac{1}{2+2\rho_p} + \frac{1}{2-2\rho_p} \geq 0, \quad \text{for} \quad \rho_p \in [-1; 1] \tag{7.3}$$

From Equations (7.1), (7.2) and (7.3) it can be observed that the OC of a 2×2 matrix is convex and symmetric in ρ_p around $\rho_p = 0$. This results in the OC being unable to discern ρ_p from $-\rho_p$. That is, negative correlation of X_1 and X_2. The $\Omega(\rho_p; X)$ as a function of ρ_p can be seen in Figure 7.1. Notice the convexity and symmetry of the OC.

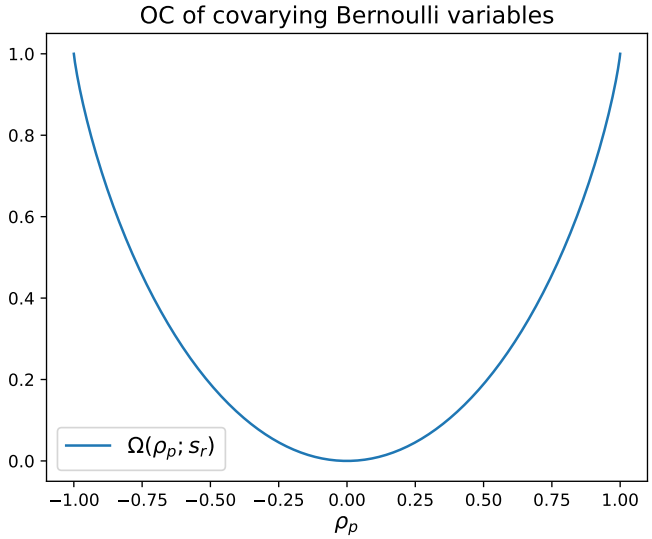


Figure 7.1: Omega complexity of two Bernoulli variables as a function of their correlation.

Example 7.5 shows that the OC does not distinguish between positive and negative covariance between two variables. This is an example of property b) in Theorem 7.4 which only specifies that the OC is maximised for dependency matrices with linearly dependent columns. In Corollary 7.6 a specific type of matrix is shown to maximise the OC even though it does not consist entirely of ones.

Corollary 7.6
Any matrix $C^{\rho,S}$ of the form

$$c_{ij}^{\rho,S} = \begin{cases} 1, & \text{for } (i+j) \bmod 2 = 0 \\ -1, & \text{for } (i+j) \bmod 2 = 1 \end{cases}$$

will result in $\Omega(\rho; S) = 1$.

Proof: Observe that all odd columns of $C^{\rho,S}$ are equal and the same is true for all even columns of $C^{\rho,S}$. From the fact that odd columns are of the form

$$\begin{bmatrix} 1 & -1 & \cdots \end{bmatrix}^T$$

and all even columns are of the form

$$\begin{bmatrix} -1 & 1 & \cdots \end{bmatrix}^T$$

it is seen that they are linearly dependent of each other with scalar -1. Thus all columns are linearly dependent, $C^{\rho,S}$ has rank 1 and it then follows from Theorem 7.4 that $\Omega(\rho; S) = 1$.

■

The following subsections seek to uncover the behaviour of the OC when treating

- variables with synergistic and redundant dependencies which are contained in more than two variables,
- matrices $C^{\rho,S}$ which can be any matrix from the set $\{C \in \mathbb{R}^{n\times n} || c_{ij}| \leq 1, c_{ii} = 1\}$, and
- specific subsets of $\mathcal{C}_{\rho,S}^n$ in Definition 7.2 which arise from specific sets S and dependency functions ρ.

7.1.1 OC of Redundant and Synergistic Variables

The OC is unsuited for detecting redundancy and synergy between more than two variables.

Proposition 7.7 (OC of Redundant and Synergistic Variables)
The omega complexity of a set of stochastic variables S does not vary with the amount of synergistic or redundant dependency found in more than two variables of S at a time.

Proof: Let $S = \{X_k\}_{k=1}^n$ and $T = \{Y_k\}_{k=1}^n$ be sets of stochastic variables, where

$$I(X_i; X_j) = I(Y_i; Y_j), \quad i, j \in \{1, \ldots, n\},$$

and

$$\begin{aligned} &\exists i, j \in \{1, \ldots, n\} \exists \mathcal{S} \subseteq S : I(X_i; X_j) \neq I(X_i; X_j | \mathcal{S} \backslash \{X_i, X_j\}), \quad i \neq j \\ &\forall i, j \in \{1, \ldots, n\} \forall \mathcal{T} \subseteq T : I(Y_i; Y_j) = I(Y_i; Y_j | \mathcal{T} \backslash \{Y_i, Y_j\}), \quad i \neq j. \end{aligned} \tag{7.4}$$

The sets S and T as such have equal pairwise redundancies from Equation 16 while S contains some synergistic or redundant dependency not found in pairwise relations. This is not the case for T. Recall that MI captures any and all dependency between two variables. Assume now for contradiction that $\Omega(I; S)$ varies with the amount of synergy or redundancy in S and that $C^{I,S}$ and $C^{I,T}$ are constructed with MI such that

$$\begin{aligned} c_{ij}^{I,S} &= I(X_i; X_j), \\ c_{ij}^{I,T} &= I(Y_i; Y_j). \end{aligned}$$

Then $C^{I,S} = C^{I,T}$ implying $\Omega(I; S) = \Omega(I; T)$ even though $S \neq T$. Therefore $\Omega(I; S)$ does not vary with the amount of synergy in S.

■

The existence of sets S and T satisfying Equation (7.4) is demonstrated by letting S be an n-bit XOR or copy and T be n independent Bernoulli variables. Proposition 7.7 shows that the OC does not capture any synergistic dependency or redundancy between more than two variables stemming from the fact that it relies upon bivariate dependency functions which only capture pairwise interaction between variables. Other methods for conditioning on other variables exits – for example partial and conditional correlation Baba et al. [2004].

7.1.2 Simulated S and C^S

Prediction of the OC of a set of more than two signals is difficult for various reasons:

- A lack of analytic descriptions of the signals prohibits further analysis.
- If an analytic description is available, the optimisation problem of the OC using analytic descriptions of the signals quickly grows inhibitively complex due to the eigendecomposition of $C^{\rho,S}$ in Definition 7.3.

- As the number of variables grows large it becomes increasingly difficult to intuitively predict how the OC is affected by changes to the variables.

Motivated by the above difficulties this section relies on simulations of sets S of variables and dependency matrices $C^{\rho,S}$ to uncover properties of the OC.

The set $\mathcal{C}^n_{\rho,S}$ in Definition 7.2 includes all dependency matrices $C^{\rho,S}$ obtainable under the conditions in Definition 7.3. This set is however difficult to analyse because of a lack of understanding of the characteristics of the matrices which it contains. The initial analysis therefore treats an easier to handle superset of $\mathcal{C}^n_{\rho,S}$.

Definition 7.8 (Superset of Dependency Matrices)
The set

$$\mathcal{C}^n = \Big\{ C \in \mathbb{R}^{n\times n} \,|\, |c_{ij}| \leq 1, c_{ii} = 1, c_{ij} = c_{ji}, i, j \in \{1, \ldots, n\} \Big\},$$

which is a superset of dependency matrices as defined in Definition 7.2 of dimension n, is referred to as the superset of dependency matrices.

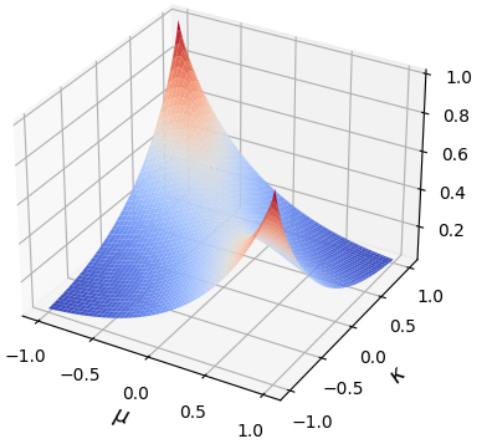
Notice, that it might be difficult to discern whether $\mathcal{C}^n_{\rho,S}$ for specific ρ and S is a true subset of $\mathcal{C}^n$.

If a finite dimension n of $\mathcal{C}^n$ is chosen, it is feasible to sample the possible combinations of the entries in $C \in \mathcal{C}^n$ and thereby the possible OCs arising from these. This is done for 3×3 matrices – that is $\mathcal{C}^3$ is examined by varying the off-diagonal entries in $C \in \mathcal{C}^3$ through the interval $[-1; 1]$. The OC then becomes dependent on three variables, i.e. the number of off-diagonal entries which can be independently varied, since C is symmetric. This is difficult to plot and therefore a three-dimensional plot is instead made where one off-diagonal entry is held constant at $\lambda \in [-1; 1]$ while the other off-diagonal entries are swept through $[-1; 1]$ such that the matrices are described by

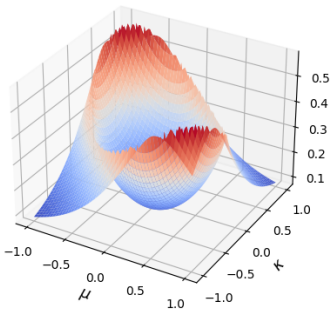
$$C = \begin{bmatrix} 1 & \lambda & \mu \\ \lambda & 1 & \kappa \\ \mu & \kappa & 1 \end{bmatrix} \quad \text{for} \quad \kappa, \mu \in [-1; 1]. \tag{7.5}$$

The OC is then seen as a function of two off-diagonal entries κ and μ, and λ can afterwards be varied. In Figure 7.2 the OC for matrices C described in Equation (7.5) are plotted with $\lambda \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$. From Figure 7.2 it can be seen that the OC is not convex in neither μ nor κ and is not consistently monotonically increasing in neither $|\mu|$ nor $|\kappa|$. This is counter-productive, as an increase in absolute values of dependency should be expected to increase total dependency as quantified by the OC.

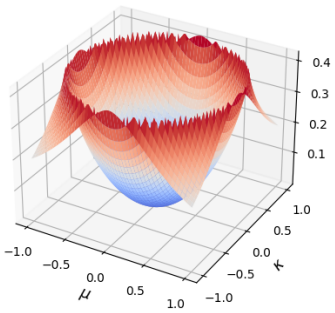
The fact that the OC is neither convex nor monotonically increasing in absolute dependency is problematic for the interpretation of the OC, since this causes the OC to confound different matrices and to not reflect an absolute increase in dependency as an increase in OC.



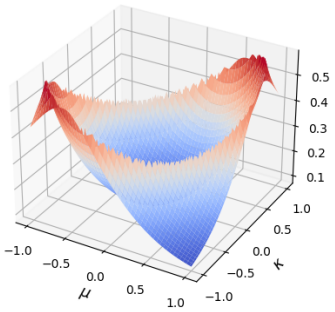
(a) $\lambda = -1$



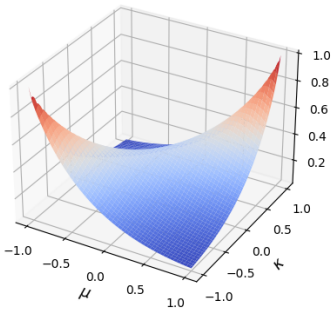
(b) $\lambda = -1/2$



(c) $\lambda = 0$



(d) $\lambda = 1/2$



(e) $\lambda = 1$

Figure 7.2: OC of the matrix C described in Equation 7.5 for different λ with $\kappa, \mu \in [-1; 1]$ with step size of $1/50$.

Consider the matrices constructed according to

$$C = \begin{bmatrix} 1 & \eta & \cdots & \eta \\ \eta & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \eta \\ \eta & \cdots & \eta & 1 \end{bmatrix}, \quad \eta \in [-1; 1], \tag{7.6}$$

where off-diagonal entries are then gradually and simultaneously increased from -1 to 1. The OC of matrices constructed as in Equation (7.6) and for dimensions $n \in \{3, 4, 5, 6\}$ can be seen in Figure 7.3.

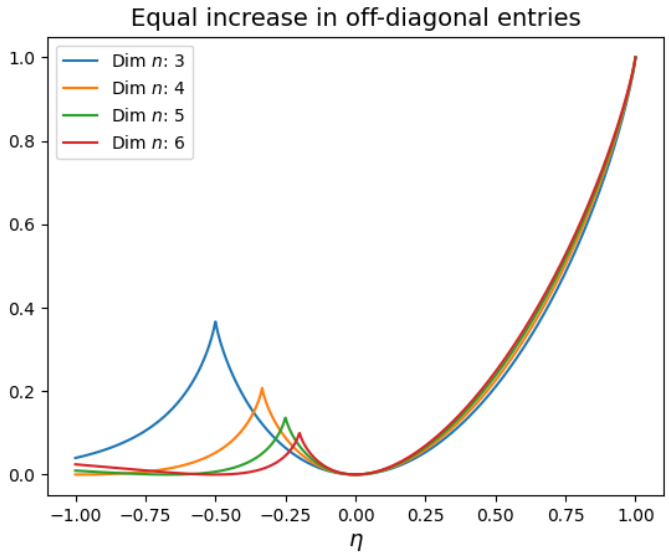


Figure 7.3: OC of C as given in Equation (7.6) with dimensions $n = \{3, 4, 5, 6\}$ and $\eta \in [-1; 1]$.

A number of observations from Figure 7.3 can be done:

- The minima in the graphs coincide with the minimum in Theorem 7.4.
- There is a local maximum with a location estimated at $\eta = -1/(n-1)$.
- For $n > 4$ there is furthermore a local maximum at $\eta = -1$.
- There is one non-trivial local minimum for $\eta < 0$.
- The OC is in general not monotonically increasing in the sum of the absolute value of off-diagonal entries. An increase may even result in a lower OC, which is counter-intuitive.
- The OC is not a convex function on $\mathcal{C}^n$.

Non-monotonicity and non-convexity are problematic since they compromise the ability of the OC to quantify dependency in a meaningful way. Sets of variables with higher absolute correlation than others might results in a lower OC and several different configurations of correlations may result in the same OC. Notice however that the simulations above treat matrices in $\mathcal{C}^n$ which are not necessarily contained in $\mathcal{C}^n_{\rho,S}$ on which the OC might be more well-behaved.

The set $\mathcal{C}^n$ includes covariance matrices which arise from common choices of ρ such as covariance, Pearson correlation coefficient and circular correlation coefficient, which are positive semi-definite and therefore a convex set. 3d plots of the same simulations as in Figure 7.2 but restricted to $\{C \in \mathcal{C}^n \,|\, C \geq 0\}$ are seen in Figure 7.4. From these figures it seems that the OC is convex in μ and κ on the set $\{C \in \mathcal{C}^n \,|\, C \geq 0\}$ for a constant λ. This can also be seen in Figure 7.5 which is analogous to Figure 7.3 but with only the cases where $C \geq 0$ are shown.

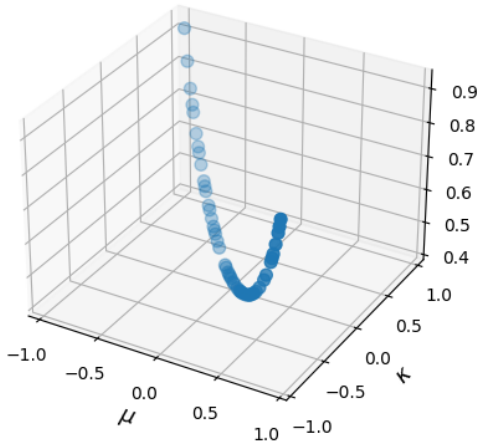
Instead of simulating a dependency matrix C directly, the following sections will examine the behaviour of the OC when applied on simulated and real variables.

Phase Shifted Sine Waves

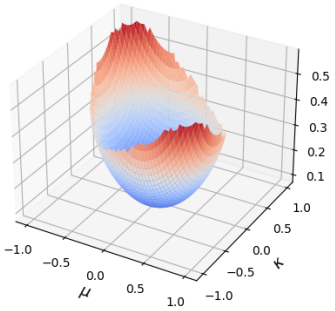
Simulated sine waves provide a simple basis for analysis of the OC on periodic signals. In Figure 7.6 is seen the OC $\Omega(\rho_p; s_1, s_2, s_3)$ of three phase shifted sine waves which are defined as

$$\begin{aligned} s_1(t) &= \sin(t), \\ s_2(t) &= \sin(t + \phi_2), \\ s_3(t) &= \sin(t + \phi_3), \end{aligned} \tag{7.7}$$

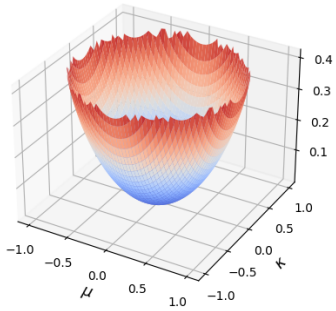
where $\phi_2, \phi_3 \in [-\pi,\pi)$ such that the phase shift of s_1 is held constant while the other two signals are phase shifted. The surface plot shows how the OC changes as a function of ϕ_2 and ϕ_3.



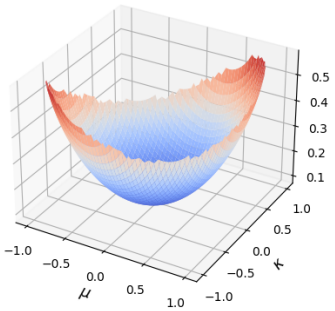
(a) $\lambda = -1$



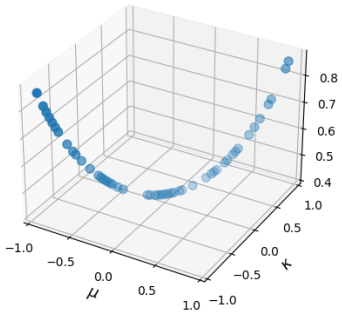
(b) $\lambda = -0.5$



(c) $\lambda = 0$



(d) $\lambda = 1/2$



(e) $\lambda = 1$

Figure 7.4: OC of the matrix C described in Equation 7.5 for different μ and $\lambda, \kappa \in [-1; 1]$ with step size of $1/25$ with only cases where $C \geq 0$.

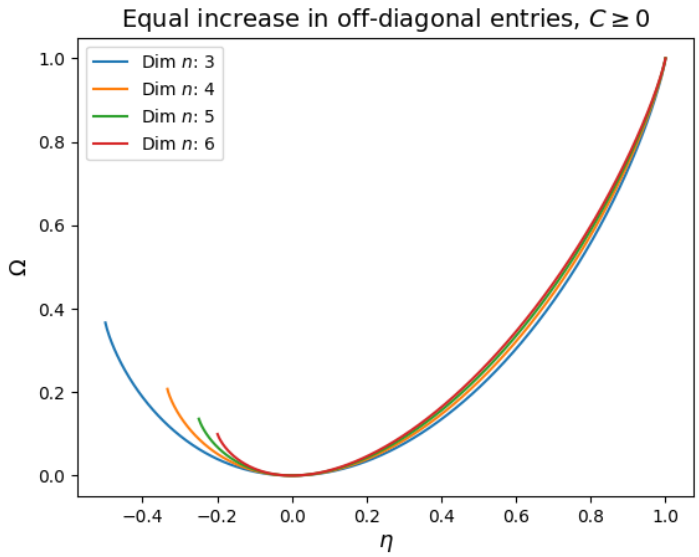


Figure 7.5: OC of C as given in Equation (7.6) with dimensions $n = \{3, 4, 5, 6\}$ and $\eta \in [-1; 1]$ with only cases where $C \geq 0$ shown.

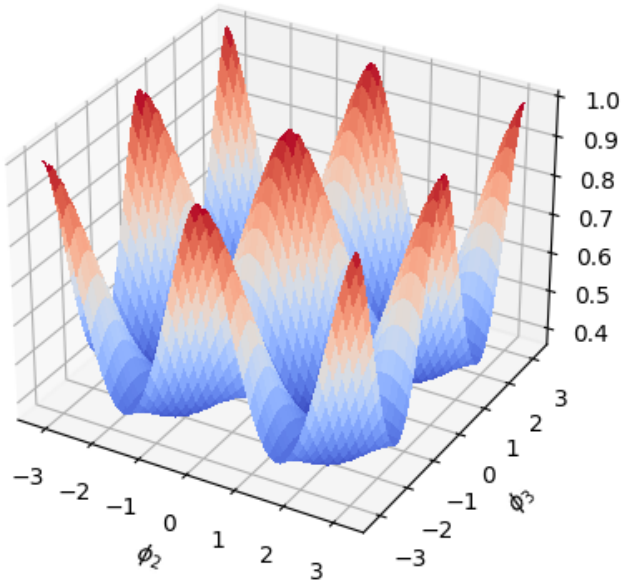


Figure 7.6: $\Omega(\rho_p; s_1,s_2,s_3)$ with the signals described in Equation (7.7) as a function of phase shifts ϕ_2 and ϕ_3 of s_2 and s_3, respectively.

Figure 7.6 shows that the OC has nine peaks which correspond to nine different situations of correlation of either 1 or -1 between all three signals. These are the exact situations where all columns are linearly dependent such that the rank of $C^{\rho_p,S}$ is 1 in accordance with b) in Theorem 7.4 and Corollary 7.6.

Four cases of $C^{\rho_p,S}$ from three phase shifted sine waves are seen in Equations (7.8)-

(7.11). See Appendix A for the equations.

$$\begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix}, \qquad \Omega(\rho_p; S_1) = 0.369070 \tag{7.8}$$

$$\begin{bmatrix} 1 & -0.30899 & -0.95105 \\ -0.30899 & 1 & -0.00006 \\ -0.95105 & -0.00006 & 1 \end{bmatrix}, \qquad \Omega(\rho_p; S_2) = 0.420613 \tag{7.9}$$

$$\begin{bmatrix} 1 & -0.30899 & -0.80899 \\ -0.30899 & 1 & -0.30909 \\ -0.80899 & -0.30909 & 1 \end{bmatrix}, \qquad \Omega(\rho_p; S_3) = 0.388521 \tag{7.10}$$

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Omega(\rho_p; S_4) = 0.420620 \tag{7.11}$$

Notice that even though Equation (7.8) displays the highest average absolute correlation between three signals (average of the absolute value of the off-diagonal entries) it results in the lowest OC. The same can be said for Equation (7.10) in which the OC is lower than the one in Equation (7.9). Finally, even though Equation (7.11) has the lowest average absolute correlation between the three signals it results in the highest OC. The main takeaway is that linear or near-linear dependency results in a high OC while the absolute value of the entries in $C^{\rho,S}$ is less indicating of high OC. This follows naturally from the OC being based upon eigendecomposition. If this phenomenon is considered to be an indication of high dependency in a set of variables, then the OC captures this well.

7.1.3 Dependency on Number of Variables

In Jia et al. [2018] it is noted that the OC is not invariant to the number of variables treated. Notice that the OC is not constant in Figures 7.3 and 7.5 for constant η and varying dimension n of $C \in \mathbb{R}^{n\times n}$ which represents the number of variables considered. This can furthermore be seen in Figure 7.7 which shows the OC of C as given in Equation (7.6) and as a function of the number of variables ranging from 3 to 64 for $\eta = 1/2$.

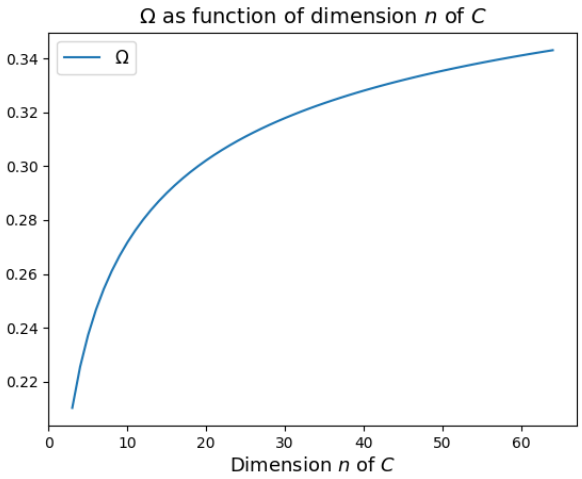


Figure 7.7: The OC as a function of dimension n of C from Equation (7.6) with $\eta = 1/2$.

With the exception of the observation by Jia et al. [2018] the analyses of the OC performed in Sections 7.1.1, 7.1.2 and 7.1.3 have as far as the authors of this project are concerned not been performed as thoroughly before and will be used as offset for the last part of the chapter where an modification of the OC is proposed.

7.2 Improving the OC

From the above analysis and examples of the OC it is clear that it has some disadvantages which are present depending on the type of variables treated and the dependency function used. The OC as defined in Definition 7.3 displays the following problems:

Problem 1 It is unable to detect synergistic and redundant dependencies in more than two variables. This is direct consequence of its restriction to bivariate dependency functions.

Problem 2 It is not in general monotonically increasing in the absolute value of the off-diagonal entries of $C^{\rho,S}$ leading to it confounding dependency matrices with each other.

Problem 3 It has non-trivial local maxima and minima which are not easily interpreted.

Problem 4 It is not invariant to the number of variables treated.

In this section a number of modifications to the OC in Definition 7.3 are presented in order to rectify Problems 1, 3 and 4 making it useful in a larger subset of $\mathcal{C}^n$ and therefore a more general setting.

7.2.1 Problem 1: Limitation of Bivariate Dependency Functions

The problem of the OC not detecting synergistic and redundant dependency between more than two variables in a set of variables can be ignored, assuming that the variables contain no such dependencies or that this type of dependency is irrelevant for the problem at hand. If these solutions are inappropriate for a specific problem, the OC can be modified by letting the dependency function ρ used for creating $C^{\rho,S}$ be multi variate, thus making it possible to detect dependencies between more than two variables.

This project presents a solution, reliant on the MI, for the case of three variables which motivates solutions for more than three variables although this becomes increasingly complex as the number of variables grows.

Proposition 7.9 (OC of Three Synergistic or Redundant Variables) Let $S = \{X_1, X_2, X_3\}$ be a set of stochastic. An OC $\Omega(I; S)$, with dependency matrix $C^{I,S}$ defined by

$$c_{ij}^{S} = I(X_i; X_j | \{X_1, X_2, X_3\} \backslash \{X_i, X_j\}),$$

is dependent on the amount of synergy and redundancy in all three variables of S.

Proof: Assume that S contains synergistic or redundant dependencies such that

$$\exists i, j \in \{1, \ldots, n\} : I(X_i; X_j) \neq I(X_i; X_j | \{X_1, X_2, X_3\} \backslash \{X_i, X_j\}). \tag{7.12}$$

Since the entries of $C^{I,S}$ are defined exactly as the RHS in Equation (7.12), any change in synergistic or redundant dependency in S implies a change in the entries of $C^{I,S}$. It follows the that OC is dependent on changes in redundant and synergistic dependencies between all three variables in S.

■

Proposition 7.9 shows an example of how to extend the OC to detect synergy or redundancy contained in more than two variables at a time, in this case three. In order to modify the OC to detect this type of behaviour between more than three variables the MI on the RHS in Equation (7.12) must condition on some $T \subseteq S \backslash \{X_i, X_k\}$ where the choice of T can be varied.

Notice however that it in Jakobsen [2013] is shown that dependency matrices based upon mutual information are not always positive semi-definite in contrast to covariance matrices. Using mutual information therefore poses a problem in this context as some of the unwanted properties of the OC are present when $C^{I,S}$ is not positive semi-definite.

7.2.2 Problems 2 and 3: Non-monotonicity and Non-trivial Local Extrema

The problem of non-monotonicity of the OC as a function of the average absolute value of the off-diagonal entries is an inherent problem stemming from the use of eigenvalues as a quantifier of coherence. This problem will not be addressed further, but it is noted that it seems to be confined to the matrices in $\mathcal{C}^n$ which are not positive semi-definite.

The non-trivial maxima and the non-monotonicity when all off-diagonal entries in $\mathbb{C}^{\rho,S}$ are increased equally are however treated. They are confined to negative average values of the off-diagonal entries and a straightforward fix is to use the absolute value of ρ confining the entries of $C^{\rho,S}$ to $[0;1]$ such that negative and positive correlations are confounded. Using a non-negative correlation function (MI for example) similarly avoids the problem of non-monotonicity while not distinguishing between negative and positive correlation.

Another simple method is to shift the entries of $C^{\rho,S}$ by 1 such that $c_{ij}^{\rho,S} = \rho + 1$. This causes two problems:

1. Complete negative dependency between all signals such that

$$c_{ij}^{S} = \begin{cases} 2 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

   results in an OC of 0. If negative dependency is to be detectable, the OC should reflect this through for example a negative value.

2. Independence between all signals such that

$$c_{ij}^{S} = \begin{cases} 2 & \text{for } i = j \\ 1 & \text{for } i \neq j, \end{cases} \tag{7.13}$$

   results in an OC greater than 0. A natural expectation is for the OC to be 0 when no dependency is present.

The above problems can be partially solved by shifting the resulting OC (which lies in the interval $[0;1]$) to the interval $[-1/2;1/2]$ by subtracting $1/2$. Since the OC however is not a linear function from $C^{\rho,S}$ to $\Omega(\rho;S)$ as seen in Figures 7.3 and 7.2, a set of independent signals with $C^{\rho,S}$ given as in Equation (7.13) will have $\Omega \neq 0$ when shifted by $-1/2$. This is seen in Figure 7.7.

7.2.3 Problem 4: Non-invariance to Number of Variables

The problem of non-invariance to number of variables is pronounced in Figure 7.7. The OC can be modified to be invariant to the number of variables in the case, where all off-diagonal entries are held constant. This becomes clearer in the next section.

7.2.4 Generalised Omega Complexity

The solution to problems 1, 3 and 4 proposed in this project is two-pronged and is seen below in a generalised definition of the OC.

Definition 7.10 (Generalised Omega Complexity)
Let $S = \{S_k\}_{k=1}^n$ be a set of stochastic variables, ρ a dependency function which satisfies either $\rho \in [-1;1]$ or $\rho \in [0;1]$ and $C^{\rho,S}$ the dependency matrix of S with respect to ρ. The generalised omega complexity of S with respect to ρ is then defined as

$$\Omega_{gen}(\rho;S) = \frac{\max(\lambda_1,\ldots,\lambda_n)-2}{n-1} - 1$$

where $\{\lambda_k\}_{k=1}^n$ are the eigenvalues of $\hat{C}^{\rho,S}$ defined by

$$\hat{c}_{ij}^{\rho,S} = c_{ij}^{\rho,S} + 1.$$

The two main changes from Definition 7.3 to Definition 7.10 are the shift of the entries in the dependency matrix $C^{\rho,S}$ and the use of the maximum eigenvalue of the shifted dependency matrix $\hat{C}^{\rho,S}$ instead of the entropy of the eigenvalues.

Theorem 7.11 (Properties of the Generalised Omega Complexity)
The generalised OC,

a) has range of $[-1;1]$ or $[0;1]$ for $\rho \in [-1;1]$ or $\rho \in [0;1]$, respectively,

b) is maximised when $\text{rank}(\hat{C}^{\rho,S}) = 1$ such that $\max(\lambda_1,\ldots,\lambda_n) = 2n$,

c) is 0 when all variables in S are independent,

d) is minimised when either $\hat{c}_{ij}^{\rho,S} = 0$ for $i \neq j$ and $\rho \in [-1;1]$ or when $\hat{c}_{ij}^{\rho,S} = 1$ for $i \neq j$ and $\rho \in [0;1]$,

e) is scale invariant if ρ is scale invariant such that $\rho(as_i,as_j) = \rho(s_i,s_j)$, for any $a \in \mathbb{R}\backslash 0$,

f) is linear in equal changes in equal off-diagonal entries in $\hat{C}^{\rho,S}$, and

g) is independent of the cardinality of S if all off-diagonal entries are equal.

Proof:

a) Since $\text{tr}(\hat{C}^{\rho,S}) = 2n$ and all entries are upper bounded by 2 the maximum possible eigenvalue is $2n$. This results in

$$\Omega_{gen}(\rho,S)(\rho;S) = \frac{2n-2}{n-1} - 1 = 1.$$

If $\rho \in [-1;1]$ then the smallest possible largest eigenvalue is 2 given by Corollary 4.5 when $\hat{C}^{\rho,S}$ is diagonal resulting in

$$\Omega_{gen}(\rho;S) = \frac{2-2}{n-1} - 1 = -1.$$

If $\rho \in [0;1]$ then the smallest possible largest eigenvalue is $n+1$ which follows from Theorem 4.1 and the decomposition of $\hat{C}^{\rho,S}$ into the identity and a matrix consisting of only ones. This yields

$$\Omega_{gen}(\rho;S) = \frac{n-1}{n-1} - 1 = 0.$$

b) If $\text{rank}(\hat{C}^{\rho,S}) = 1$ then $\text{null}(\hat{C}^{\rho,S}) = n-1$ which results in an eigenvalue of 0 with multiplicity $n-1$ and therefore only one non-zero eigenvalue which is equal to $\text{tr}(\hat{C}^{\rho,S}) = 2n$.

c) If all variables in S are independent then $\hat{c}_{ij}^{\rho,S} = 1$ for $i \neq j$, which from Theorem 4.7 results in a largest eigenvalue of $n+1$ yielding

$$\Omega_{gen}(\rho;S) = \frac{n-1}{n-1} - 1 = 0.$$

d) For the case where $\rho \in [0;1]$ the proof is given by a) and c). When $\rho \in [-1;1]$ the OC is minimised when all eigenvalues are equal. This is achieved when all off-diagonal entries in $\hat{C}^{\rho,S}$ are 0, such that the largest eigenvalues is 2, which follows from the fact that $\text{tr}(\hat{C}^{\rho,S}) = 2n$. The OC is then $(2-2)/(n-1)-1 = -1$.

e) If ρ is scale invariant, then $\Omega_{gen}(\rho;S)$ is trivially scale invariant.

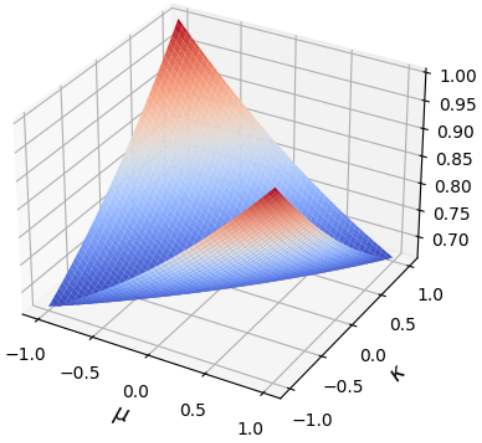
f) Linearity of the eigenvalues of $\hat{C}^{\rho,S}$ in equal changes in the off-diagonal entries in $\hat{C}^{\rho,S}$ follows from Theorem 4.7, and

g) Follows trivially from Theorem 4.7.

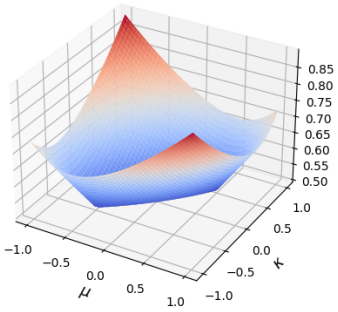
■

The generalised OC has additional properties which are not true for the OC and these properties show that it is useful in a more general sense.

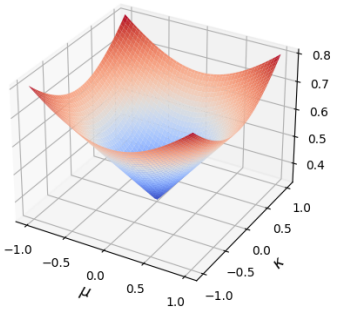
3d plots analogous to those in Figure 7.2 are seen in Figure 7.8.



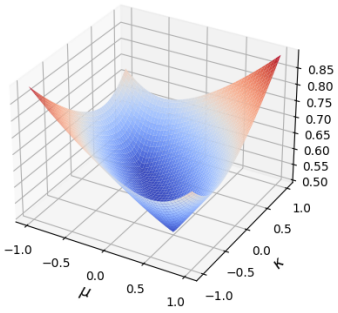
(a) $\mu = -1$



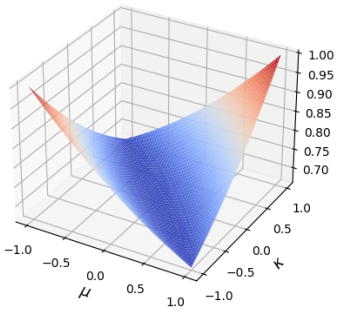
(b) $\mu = -1/2$



(c) $\mu = 0$



(d) $\mu = 1/2$



(e) $\mu = 1$

Figure 7.8: Generalised OC of the matrix C described in Equation 7.5 for different λ and $\mu, \kappa \in [-1; 1]$ with step size of $1/25$.

Notice in Figure 7.8 that the generalised OC seems to be monotonic in κ and

μ for constant λ and also convex in those arguments. This is the case even though indefinite matrices are considered, which might allow use of for example MI as dependency function without the problems of the OC.

The linearity of the OC in equal increases in equal off-diagonal entries and invariance to the number of variables is furthermore seen in Figure 7.9.

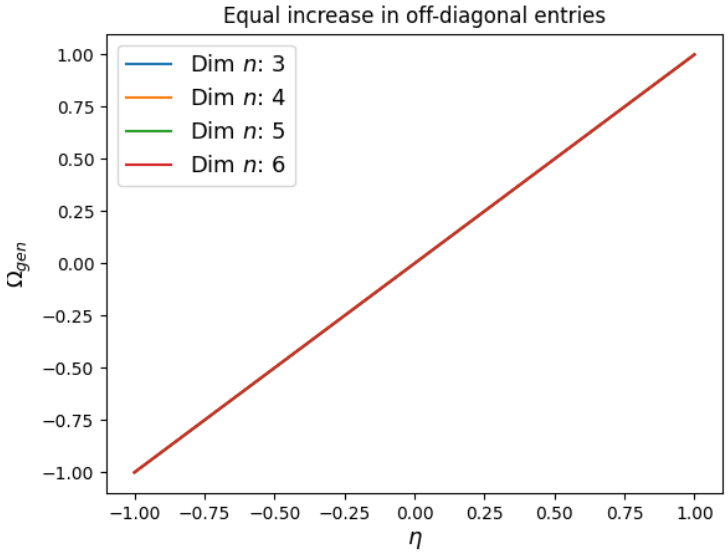


Figure 7.9: Generalised OC of C as given in Equation (7.6) with dimensions $n = \{3, 4, 5, 6\}$ and $\eta \in [-1; 1]$.

7.3 Summary

The analysis performed and examples given in this chapter established some properties of the OC and uncovered a number of problems, which seem to be unaccounted for in existing literature.

These problems where noteworthy in the subset $\mathcal{C}^n$ described in Definition 7.2 but some of them seem of limited influence in the subset of positive semi-definite matrices.

A generalised omega complexity which allows for better distinction between negative and positive dependencies and which behaves in a more consistent way as a function of dependency between signals was presented. It is proposed to future users of the omega complexity to consider if the generalised version in Definition 7.10 is more suitable for their problem and furthermore to show caution and consider what type of signals they are applying the mathematical tool to since the properties of $\Omega_{gen}(\rho, S)$ on $\mathcal{C}^n_{\rho,S}$ are not fully accounted for.

8 | Graph Theory

This chapter serves as an introduction to notions regarding graph theory. It is introduced as a method of analysing connectivity in larger networks, in this project for EEG analysis. The chapter is based loosely on Diestel [2017], Bondy and Murty [2008], and Bondy and Murty [1982].

8.1 Fundamentals

Informally, a graph is a mathematical object consisting of nodes connected by edges. A simple graph with 3 nodes, and 1 undirected edge can be seen in Figure 8.1. The terms directed and undirected edge will be elaborated upon later in the chapter.

Definition 8.1 (Graph and Incidence)
Let N be a set of nodes, E be a set of edges and P be set of pairs of nodes from N. An incidence function is a map $\mathcal{I} : E \to P$ such that $\forall e \in E$, $\exists a,b \in N : \mathcal{I}(e) = e_{ab}$, and e_{ab} denotes the pair a and b connected by an edge. Then $G = (N, E, \mathcal{I})$ is called a graph.

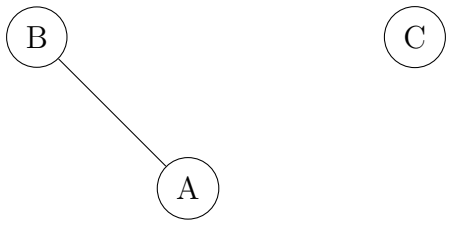


Figure 8.1: An example of a graph.

Definition 8.2 (Node and Edge Set Funtions)
Let $G = (N, E, \mathcal{I})$ be given. The node and edge set functions of G are

$$\mathcal{N}(G) = N \quad \text{and} \quad \mathcal{E}(G) = E,$$

respectively.

Note that $|\mathcal{N}(G)| = |N|$ denotes the number of elements in the node set and $|\mathcal{E}(G)| = |E|$ denotes the number of elements in the edge set.

Graph objects can be classified into two main groups namely directed and undirected graphs.

Definition 8.3 (Directed and Undirected Graph)
Let $G = (N, E, \mathcal{I})$ be a graph, with $\mathcal{I} : E \to P$. If $P = N \times N$ then G is called a directed graph. If $P = \{\{a,b\}|a,b \in N\}$ then G is called an undirected graph.

Definition 8.4 (Weight Map)
Let $G = (N, E, \mathcal{I})$ be a graph and $\mathcal{I} : E \to P$. A weight map of G is a map $W_G : P \to \mathbb{R}^+$ such that

$$W_G(x, y) = \begin{cases} w_{xy}, & e_{xy} \in P \\ 0, & \text{otherwise,} \end{cases}$$

where $w_{xy} \geq 0$ is referred to as the weight of the edge e_{xy}.

This motivates talking about weighted and unweighted graphs.

Definition 8.5 (Weighted and Unweighted Graphs)
Let $G = (N, E, \mathcal{I})$ be a graph and $\mathcal{I} : E \to P$ be given. G is called unweighted if $w_{xy} = 1$ for all $e_{xy} \in P$, otherwise G is called weighted.

In order to formalise the weights, and draw connections to linear algebra, the adjacency matrix is introduced.

Definition 8.6 (Adjacency Matrix)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G, and $N = \{n_0, \ldots, n_{|N|-1}\}$ be given. The matrix $A \in \mathbb{R}^{|N|\times|N|}$ with entries

$$A_{ij} = W_G(n_i,n_j)$$

is called the adjacency matrix of G.

The above definitions are used as offset for more advanced tools for analysis of graphs.

8.2 Clustering

Consider that certain subsets of nodes and edges can form groups with high internal connectivity and lower connectivity between groups. Such groups can be described through clusterings. This section will describe how to define them, and evaluate them, and is based on Almerida et al. [2019] and Newman and Girvan [2003].

Definition 8.7 (Unions and Intersections of Graphs)
Let $G = (N_1, E_1, \mathcal{I})$ and $H = (N_2, E_2, \mathcal{I})$ be graphs. The union of G and H is defined as

$$G \cup H = (N_1 \cup N_2, E_1 \cup E_2, \mathcal{I}),$$

and the intersection is defined as

$$G \cap H = (N_1 \cap N_2, E_1 \cap E_2, \mathcal{I}).$$

Definition 8.8 (Cluster and Clustering)
Let $G = (N, E, \mathcal{I})$ be a graph and $N_i \subseteq N$ such that $N = \bigcup_{i=1}^{n} N_i$ and $N_i \bigcap N_j = \varnothing$ for all i, j. Furthermore let

$$E_i = \{e \in E | \mathcal{I}(e) = e_{ab}, \quad a, b \in N_i\},$$

such that $C_i = (N_i, E_i, \mathcal{I})$ is a graph. Then $C = \bigcup_{i=1}^{n} C_i$ is called a clustering, and C_i is called a cluster.

Note that Definition 8.8 does not consider the performance of a clustering, and it is hence not given that a clustering will reflect underlying communities in a meaningful way.

8.2.1 Evaluations of Clusterings

In this section methods for evaluating whether or not a clustering reflects underlying communities will be introduced.

Consider that one interesting property of a community is that a subset of the nodes does not have a high degree of connectivity with nodes not in that community.

Definition 8.9 (Intercluster Weight Sum)
Let $G = (N, E, \mathcal{I})$ be a graph with clustering $C = \bigcup_{i=1}^{n} C_i$ and corresponding weight map W_G. The intercluster weight sum of C_i is defined as

$$ICWS(i) = \sum_{a \in \mathcal{N}(C_i)} \sum_{b \in N \setminus \mathcal{N}(C_i)} W_G(a, b) + W_G(b, a), \quad i = 1, \ldots, n.$$

In order to quantify the connectivity internally versus externally in a cluster, cluster degree is introduced.

Definition 8.10 (Cluster and Non-Cluster Degree)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G and let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G. The cluster degree of C_i is

$$CD(i) = \sum_{a \in \mathcal{N}(C_i)} \sum_{b \in N} W_G(a, b) + \sum_{a \in N \setminus \mathcal{N}(C_i)} \sum_{b \in \mathcal{N}(C_i)} W_G(a, b).$$

The non-cluster degree of C_i is

$$NCD(i) = \sum_{a,b \in N \setminus \mathcal{N}(C_i)} W_G(a, b).$$

Note that the cluster degree captures information about how much weight is related to a cluster, while the non-cluster degree quantifies how much weight is not associated with a given cluster.

A characteristic of a good clustering, is that the connectivity between clusters is relatively low compared to the connectivity internally in the clusters.

Definition 8.11 (Conductance)
Let $G = (N, E, \mathcal{I})$ be a graph with W_G as a weight map. Furthermore let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G. The conductance of cluster C_i is then

$$CON(i) = \begin{cases} 0, & \text{for } \mathcal{N}(C_i) \notin \{\varnothing, N\} \text{ and } ICWS(i) = 0 \\ 1, & \text{for } \mathcal{N}(C_i) \in \{\varnothing, N\} \\ \frac{ICWS(i)}{\min\{CD(i), NCD(i)\}}, & \text{otherwise.} \end{cases}$$

Conductance is defined for a single cluster, and serves to balance the relationship between weights related to the cluster and the weights not related to the cluster.

To evaluate the performance of a clustering, the conductance is generalized from a single cluster to a clustering.

Definition 8.12 (Intercluster Conductance)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G. Furthermore let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G.

$$ICC(C) = 1 - \max_{i \in \{1,\ldots,n\}} (\mathrm{CON}(i))$$

is called the intercluster conductance of the clustering C.

Note that when constructing a good clustering it is not desirable for any cluster to have high connectivity to any other set of clusters, and hence a clustering with high intercluster conductance is desirable. A low intercluster conductance might indicate a too fine clustering. Note that the intercluster conductance is not great for distinguishing if a clustering is too coarse in comparison to any underlying communities.

Definition 8.13 (Weight Sum and Intracluster Weight Sum)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G. Furthermore let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G. Then

$$m_G = \sum_{a,b \in N} W_G(a, b)$$

is called the weight sum of G, and

$$m_I(i) = \sum_{a,b \in \mathcal{N}(C_i)} W_G(a, b)$$

is called the intracluster weight sum of C_i.

While the weight sum aggregates all weights in the graph, the intracluster weight sum quantifies the weight exclusively related to a specific cluster C_i. Based on the weight sum and intracluster weight sum the coverage is defined.

Definition 8.14 (Clustering Coverage)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G. Furthermore let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G. The coverage of C is then

$$\mathrm{COV}(C) = \frac{\sum_{i=1}^{n} m_I(i)}{m_G}.$$

Note that the coverage describes how much of much of the connectivity is contained withing the clusters relative to the connectivity of the entire graph. A clustering coverage close to 1 is desired, but does not take the internal connectivity of the clustering into consideration.

Next the concept of modularity is introduced.

Definition 8.15 (Modularity)
Let $G = (N, E, \mathcal{I})$ be a graph with weight map W_G. Furthermore let $C = \bigcup_{i=1}^{n} C_i$ be a clustering of G. The modularity of C is then

$$\mathrm{MOD}(C) = \mathrm{COV}(C) - \frac{1}{m_G^2} \sum_{i=1}^{n} (\mathrm{CD(i)})^2 .$$

Notice that a high modularity implies a high cluster coverage, and a low cluster degree. Hence modularity can be utilised to avoid too fine of a clustering. The remaining part of this chapter will introduce an algorithm for designating clusters of a graph.

8.3 Clustering Algorithm

As alluded to, an approach to detecting communities will be introduced. The Louvain clustering algorithm has in this project been chosen as the the method for finding clusters.

8.3.1 Louvain clustering Algorithm

This section will introduce the Louvain clustering algorithm, which based on the modularity seeks to obtain a clustering representative of the underlying communities.

Algorithm 1 Louvain Clustering

Input $G = (N, E, \mathcal{I})$, W_G
Output $C = \bigcup_{i=1}^{n} C_i$

Initialize:
each $u_i \in N$ is assigned an individual cluster C_i for $i = 1, \ldots, |N| - 1$;
set $MOD(C) = 0$ and make a list C containing all C_i;
while one or more nodes change assigned cluster **do**
    for $u_i \in N$ **do** ▷ Finds the best cluster C_j for each node u_i
        make C_N a list of clusters with at least one node adjacent to u_i;
        for $C_j \in C_N$ **do**
            test u_i to C_j and calculate $\Delta MOD(C)$;
            if $u_i \in \mathcal{N}(C_j)$ yields $\Delta MOD(C) > 0$ **then**
                keep u_i in C_j;
            else keep u_i in original cluster;
            end if
        end for
        if C contains empty clusters, remove them;
    end for
end while

As is clear from the algorithm, the approach is greedy and hence does not necessarily find an optimal clustering with regard to the modularity. It should be noted that the Louvain algorithm is dependent on the ordering of the nodes and hence the algorithm is often implemented with a random starting point.

8.4 Combined Clustering

The following steps are used in the creation of a combined clustering between two data sets:

Algorithm 2 Combined Clustering

Input $G_0 = (N_0, E_0, \mathcal{I}_0)$, $G_1 = (N_1, E_1, \mathcal{I}_1)$
Output $C^r = \bigcup_{k=0}^{n} C_k^r$

1: **Initialize**:
2: $C = \{\}$
3: Apply Louvain to G_0 and G_1 resulting in $C^0 = \bigcup_{i=0}^{n} C_i^0$, and $C^1 = \bigcup_{j=0}^{m} C_j^1$
4: Let C^0 be the clustering with most clusters such that $n \geq m$
5: **for** $k = \text{range}(0, n)$ **do**
6: $\quad l = \underset{l}{\text{argmax}}\{|\mathcal{N}(C_k^0 \bigcap C_l^1)|\}))$ ▷ Finds the cluster with most shared nodes
7: $\quad N_r = \mathcal{N}(C_k^0 \cap C_l^1)$
8: $\quad E_r = \mathcal{E}(C_k^0 \cap C_l^1)$
9: $\quad$ Append $C_k^r = (N_r, E_r, \mathcal{I})$ to C
10: **end for**

9 | Simulated signals

Before testing tools for quantifying coherence in multivariate signals on real EEG signals they are tested on more controlled simulated signals. Often used ways of simulating signals mimicking the behaviour of EEG signals are the so called Rössler system and the multivariate autoregressive model.

9.1 The Rössler System

Rössler systems are used in Baboukani et al. [2018] to simulate signals which are then evaluated through phase similarity. In Subramaniyam and Hyttinen [2014] they are also used for a preliminary analysis before analysing structural properties of EEG signals. Before presenting the coupled system of individual Rössler systems, the differential equations defining the single Rössler system are presented.

Definition 9.1 (Rössler System)
The Rössler system is the solution to the set of differential equations given by

$$\begin{aligned}\frac{\mathrm{d}X}{\mathrm{d}t} &= -\omega Y - Z + \sigma\delta,\\ \frac{\mathrm{d}Y}{\mathrm{d}t} &= \omega X + aY,\\ \frac{\mathrm{d}Z}{\mathrm{d}t} &= b + Z(X - c),\end{aligned}$$

where X, Y and Z are dependent on the time variable t, the constants $a, b, c, \omega \in \mathbb{R}$ determine the behaviour of the solutions and $\delta_j \sim \mathcal{N}(0,1)$ is Gaussian noise weighted by the constant σ.

The Rössler system in Definition 9.1 consists of three coupled signals X, Y and Z whose characteristics can be altered by changing the constants in the system. The effect of changing these constants will not be treated in this project and only the time series X is considered as output from the Rössler system.

9.2 Coupled Rössler Systems

As the goal of the chapter is to simulate a signal which is similar in nature to EEG signals more than one signal is needed. In this section a method of generalising the Rössler system to a coupling of Rössler systems is introduced. A method of generalising to n Rössler systems is defined below.

Definition 9.2 (Coupled Rössler Systems)
Let a system be described by

$$\begin{aligned}
\frac{\mathrm{d}X_j}{\mathrm{d}t} &= -\omega_j Y_j - Z_j + \left[\sum_{i\neq j} \epsilon_{ij}(X_i - X_j)\right] + \sigma\delta_j \\
\frac{\mathrm{d}Y_j}{\mathrm{d}t} &= \omega_j X_j + aY_j \\
\frac{\mathrm{d}Z_j}{\mathrm{d}t} &= b + Z_j(X_j - c),
\end{aligned}$$

where X, Y and Z are dependent on the time variable t, the constants $a, b, c, \omega \in \mathbb{R}$ determine the behaviour of the solutions, $\delta_j \sim \mathcal{N}(0,1)$ is Gaussian noise weighted by the constant σ, $j \in \{1, \ldots, n\}$, $n \in \mathbb{N}$ and $\epsilon \in \mathbb{R}^{n\times n}$ is a matrix with coupling coefficients $\epsilon_{ij} \in \mathcal{R}$. This system is referred to as n coupled Rössler systems.

Recall that only the X time series is used as output from the Rössler systems, which implies that n coupled Rössler systems as described in Definition 9.2 outputs n time series. This project will use six coupled Rössler systems exclusively.

The main difference from Definition 9.1 to Definition 9.2 is the addition of the term $\sum_{i\neq j} \epsilon_{ij}(X_i - X_j)$, which introduces the coupling between the six X equations. This coupling is controlled through the coupling matrix ϵ whose entries in this project are either 0 or a constant η. This determines the strength of the coupling between the X terms.

The coupling of the time series X_j of n coupled Rössler systems can be visually represented through a graph such that a node j represents X_j and an edge between nodes j and k represents a coupling $\epsilon_{jk} > 0$. In Figure 9.1 can be seen six examples of graphical representations of the six coupled Rössler systems. These configurations are for the remainder of the project referred to as the six cases of couplings for the Rössler systems.

9.3 Simulation

The coupled Rössler systems in Definition 9.2 are solved numerically in Python and this is for the rest of the project referred to as simulations of coupled Rössler systems.

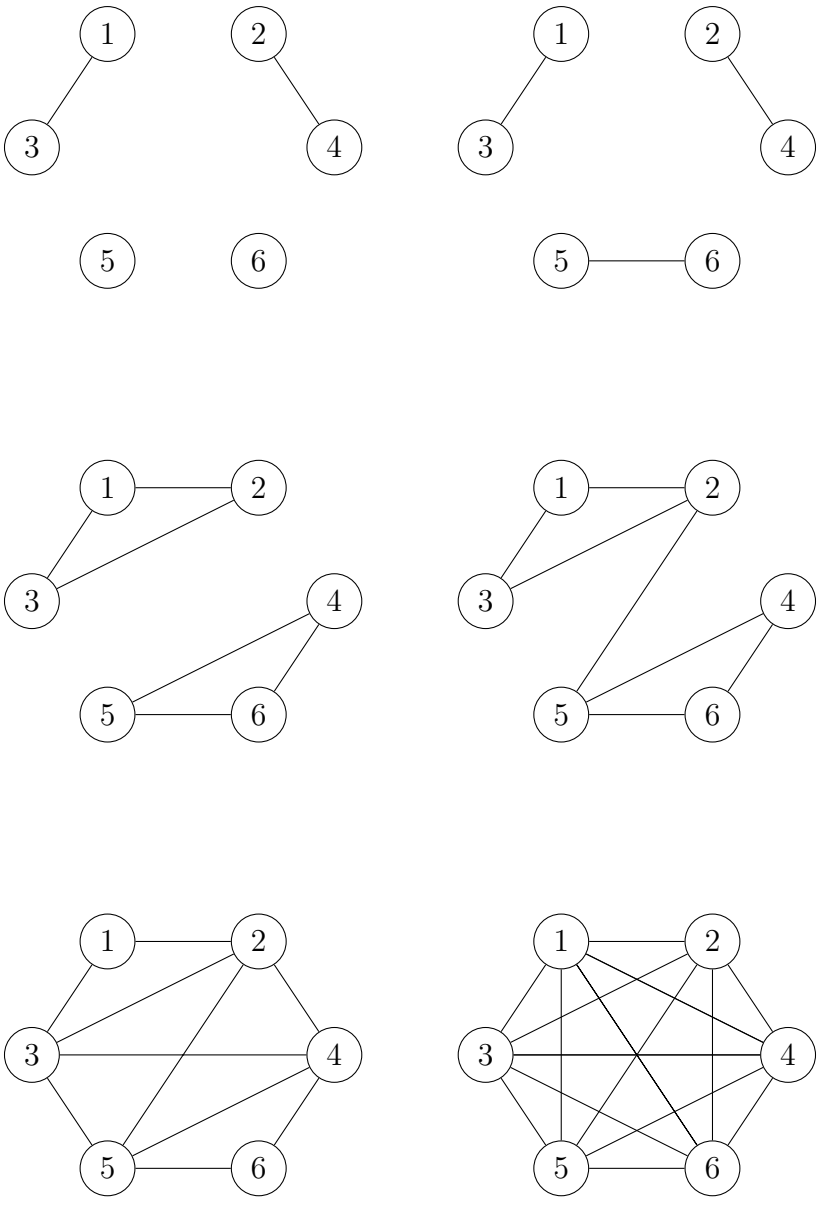


Figure 9.1: Visualisations of the six coupled Rössler systems used in the project showing whether a coupling is present through edges between nodes $1, \ldots, 6$ which represent X_j, $j \in \{1, \ldots, 6\}$.

The constant parameters which are used for all simulations of the coupled Rössler systems throughout the project are as follows:

- $a = 0.165$, $b = 0.2$, $c = 10$.
- $\omega_1 = 0.95$, $\omega_2 = 0.97$, $\omega_3 = 0.99$, $\omega_4 = 1.01$, $\omega_5 = 1.03$, $\omega_6 = 1.05$.
- $\sigma = 1.5$.

The above parameters are found in or inspired by Pounder and Sauer [2009]. The initial conditions $X_j(0) = 1$, $Y_j(0) = 1$ and $Z_j(0) = 0$ for $j \in \{1, \ldots, 6\}$ furthermore supply the same basis for all simulations whereafter they deviate as a consequence of the noise term.

The following parameters are furthermore used in the specification of the simulations and will be specified when specific simulations are produced.

- T: the time period in which the simulations are run – unless otherwise stated, this implies a time interval for the simulations such that $t \in [0; T]$.
- N: number of samples in the time period defined from T such that the sampling frequency is N/T.
- N_{disc}: discarded samples starting from $t = 0$ resulting in $t \in [N_{disc}T/N; T]$. This is to avoid the transient behaviour in the beginning of the simulations and allows them to deviate from the initial conditions.

The simulations of the six cases of six coupled Rössler systems with couplings as represented in Figure 9.1 for $\eta = 0.1$, $T = 100$, $N = 1000$ and $N_{disc} = 0$ can be seen in Figure 9.2. Notice the increased similarity between the time series as the number of couplings increase (and η is held constant) and that the time series show a period of transient behaviour in the beginning while they deviate from their initial conditions. The samples in which the transient behaviour is found will be discarded in simulation in the rest of the project.

9.4 Multivariate Autoregressive Process

In this section simulation of a specific multivariate autoregressive (MVAR) process is briefly described. In Baboukani et al. [2020] an MVAR process is used for simulating signals which have directed dependencies and AR-processes have previously been used for EEG analysis in Atyabi et al. [2016]; Zhang et al. [2018a,b]

The MVAR model consists of multiple coupled autoregressive processes each producing a time series.

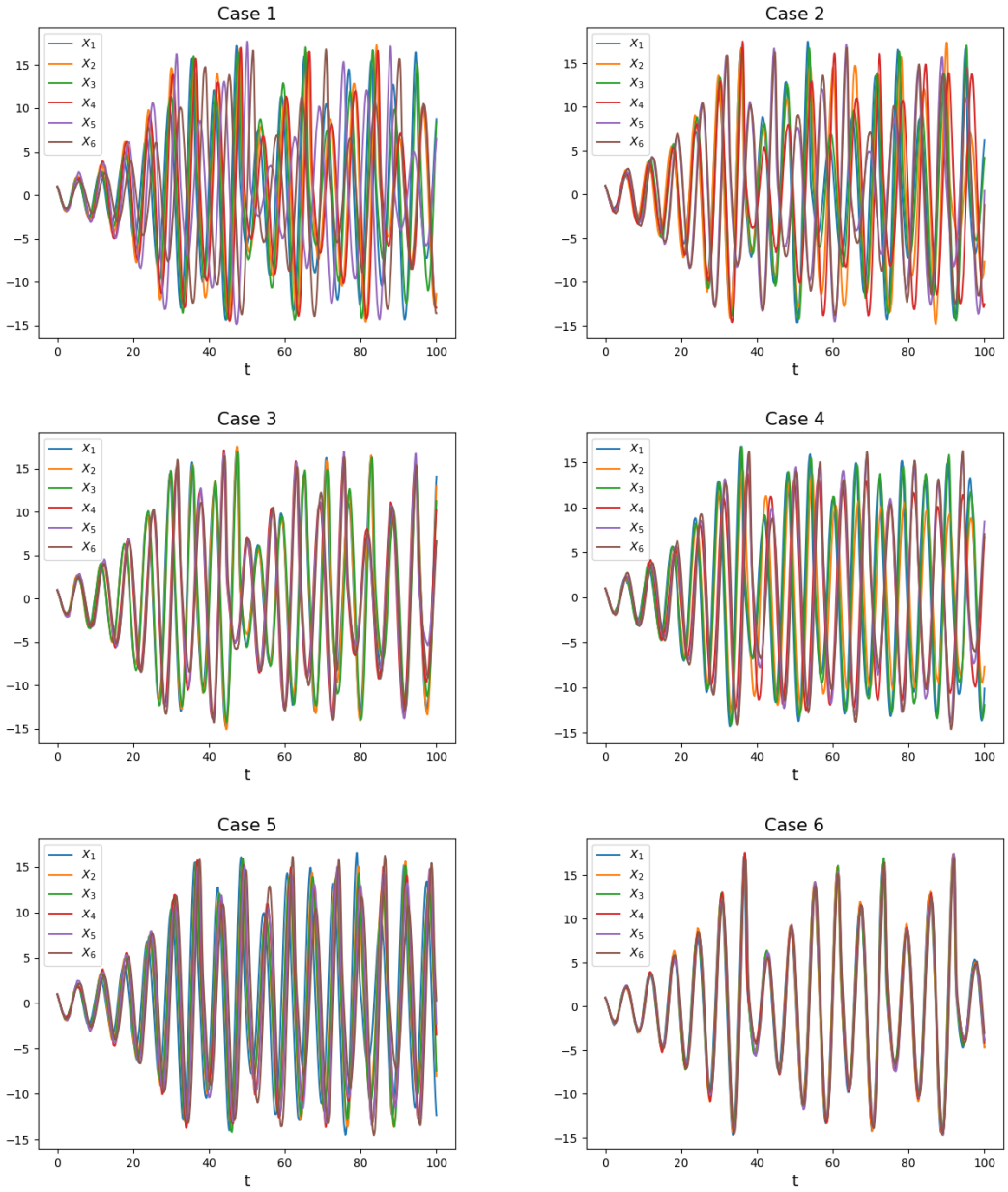


Figure 9.2: Simulations of coupled Rössler systems with $\eta = 0.1$, $T = 100$, $N = 1000$ and $N_{disc} = 0$ for each of the six different cases illustrated in Figure 9.1.

Definition 9.3 (Multivariate Autogregressive Process)
The MVAR process of order p with m channels is of the form

$$y_i(n) = -\sum_{i=1}^{p} a_{ik} y_i(n-k) - \sum_{\substack{j=1 \\ j\neq i}}^{m} \sum_{k=1}^{p} a_{jk} y_j(n-k) + \epsilon_i(n).$$

where y_i and ϵ_i are the output and a Gaussian noise input, respectively, for process i.

As such, the output of each process is dependent on the previous values of any number of the processes in the system. This causes directed spatial and temporal couplings.

The MVAR model introduced in Baboukani et al. [2020] consists of five coupled AR-processes of order 3. The dependencies and coefficients a_{ik} are as follows:

$$\begin{aligned}
y_{1,n} &= 0.95\sqrt{2}y_{1,n-1} - 0.9125y_{1,n-2} + \epsilon_{1,n} \\
y_{2,n} &= 0.5y_{1,n-2}^2 + \epsilon_{2,n} \\
y_{3,n} &= -0.4y_{1,n-3} + 0.4y_{2,n-1} + \epsilon_{3,n} \\
y_{4,n} &= -0.5y_{1,n-1}^2 + 0.25\sqrt{2}y_{4,n-1} + \epsilon_{4,n} \\
y_{5,n} &= -0.25\sqrt{2}y_{4,n-1} + 0.25\sqrt{2}y_{5,n} + \epsilon_{5,n}
\end{aligned} \tag{9.1}$$

where $\epsilon_n = [\epsilon_{1,n}, \ldots, \epsilon_{5,n}]^T \sim \mathcal{N}(0, I)$, $I \in \mathbb{R}^{n\times n}$. A simulation of the model in Equation (9.1) can be seen in Figure 9.3. The MVAR process can furthermore be altered through

$$Y_{mixed} = AY,$$

where $Y \in \mathbb{R}^{5\times N}$ is a matrix containing the MVAR process in Equation (9.1) simulated for N samples, and

$$A = \begin{bmatrix} 1-\alpha & \alpha & \cdots & \alpha \\ \alpha & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \alpha \\ \alpha & \cdots & \alpha & 1-\alpha \end{bmatrix}, \quad \alpha \in [0; 1/2]$$

is a mixing matrix which mixes the time series in Y depending on the mixing coefficient α, with $\alpha = 0$ resulting in no mixing $\alpha = 1/2$ resulting in five identical time series.

The MVAR model in Equation (9.1) is in conjunction with a mixing matrix in this project used for simulating synthetic signals with directed dependencies.

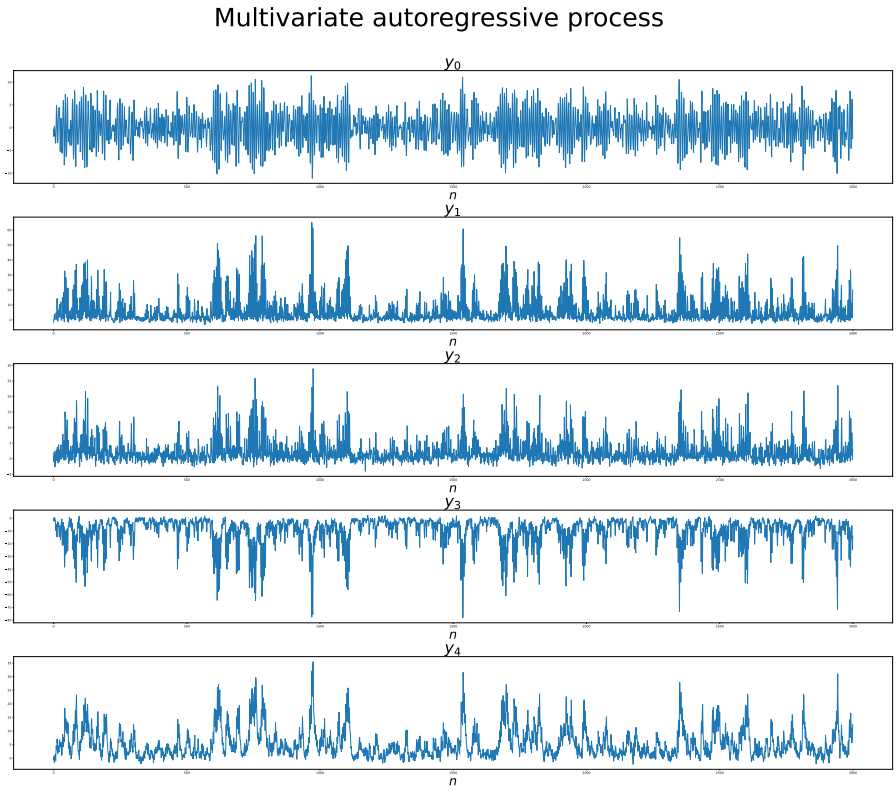


Figure 9.3: Plots of the MVAR process from Equation (9.1) simulated for $n \in \{0, \ldots, 3000\}$.

10 | Simulated Signals and Results

This chapter presents the results of applying tools from Chapters 2, 3, 5 and 7 on simulated signals in order to evaluate whether the tools are able to detect changes in dependency between signals with similar structures.

10.1 Simulated Signals

Two different models are used for simulating signals – the coupled Rössler systems from Definition 9.2 and the MVAR process from Equation 9.1.

The nature and degree of the dependencies in the simulated signals can to some degree be controlled, which allows for a better understanding of the tool applied. The coupling configurations and coefficient of the coupled Rössler systems and the mixing coefficient of the MVAR processes can be controlled and therefore used as references for the tools applied.

The coupled Rössler systems are simulated according to the six cases presented in Figure 9.1. Each of these cases are simulated with coupling coefficient $\eta \in \{0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$ resulting in 36 different configurations. These configurations are simulated 200 times with $T = 130$, $N = 7800$ and $N_{disc} = 1800$ resulting in a sample rate of 60 samples per t and the signals therefore run for $t \in [30; 130]$.

The MVAR processes are simulated with $N = 3000$ and $\alpha \in \{0, 0.1225, 0.245, 0.3675, 0.49\}$, resulting in five different configurations. Each of these are simulated 200 times.

The simulated signals will be represented in three different ways: the real signal, the instantaneous amplitude and the instantaneous phase as given through the Hilbert transform from Definition 2.4. These are denoted s_r, s_a and s_p respectively. When TE is applied on the simulated signals it is only on s_r.

10.2 Dependency Measures and Signal Representations

The tools listed below in conjunction with the listed representations are used for quantifying dependencies in the simulated signals:

- Generalised OC from Definition 7.10 with various bivariate dependency functions:
  - Sample Pearson correlation coefficient ρ_p of s_r, s_a and s_p.
  - Sample circular correlation coefficient from Equation 3.1 of ρ_c of s_p.
  - KSG estimate of mutual information $\hat{I}_{KSG,k}$ from Definition 6.8 of s_r, s_a, and s_p.
- Total correlation estimated with the entropy estimate in Definition 6.2 with $k = 4$ of s_r, s_a, and s_p.
- Dual total correlation estimated with the entropy estimate in Definition 6.2 with $k = 4$ of s_r, s_a, and s_p.
- O-information from Definition 5.28 and given by the above estimates of TC and DTC of s_r, s_a, and s_p.
- Transfer entropy from Definition 5.16 on s_r estimated with the `Bteknn` function from the `ITS` package in MATLAB which uses a kNN estimator. Here $k = 4$ and embedding delay with parameter $M = 5$ are used. As there is a difference in the way the two signals are constructed, the time lag used varies. For the MVAR processes a lag of $L = 3$ is used, since the signals at any given time only depend on the samples at the last three time stamps. For the coupled Rössler systems the time lag is $L = 5$.

For the remainder of the project, the above quantifiers of dependency are referred to as dependency measures even though they are not measures in a mathematical sense.

The discussion of these choices can be found in Section 11.2.

10.3 Clustering Method

Using the TE results as weights a graph is constructed with the intention of using the Louvain clustering algorithm to uncover the potential underlying communities present in the MVAR process. The discussion of this method is done in Section 11.4.

The clustering of the MVAR data serves as a proof of concept before applying the same method to EEG data, where the clustering will serve as a method of dimensionality reduction.

10.4 Results

The results for coupled Rössler systems and MVAR processes are presented in separate sections below. All dependency measures and signal representations with the exception of TE are used in Sections 10.4.1 and 10.4.2. The results from TE are presented in Section 10.4.3 and used for clustering in Section 10.4.4.

10.4.1 Coupled Rössler Systems

The means of the combinations of dependency measures and signal representations described in Section 10.2 of the simulated coupled Rössler systems are seen in Figure 10.1. Each figure shows how the mean of a specific type of dependency measure varies with the coupling coefficient η for a specific case of coupling.

The dual total correlation against the total correlation of the Rössler systems for the six different cases and six different ηs can be seen in Figure 10.2.

10.4.2 MVAR Processes

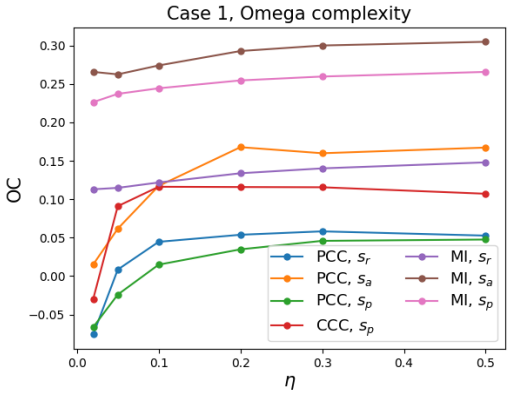
The means of the combinations of dependency measures and the different representations described in Section 10.2 of the simulated MVAR processes are seen in Figure 10.3. Each figure shows how the mean of a specific type of dependency measure varies with the mixing coefficient α.

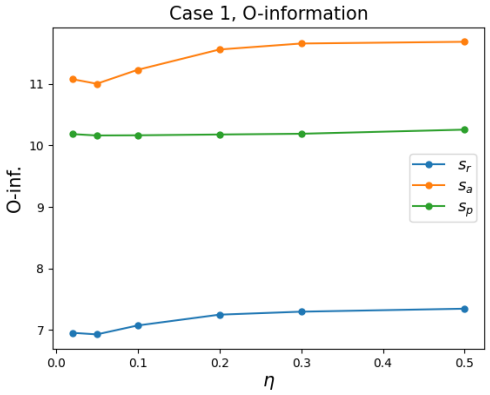
10.4.3 Transfer entropy

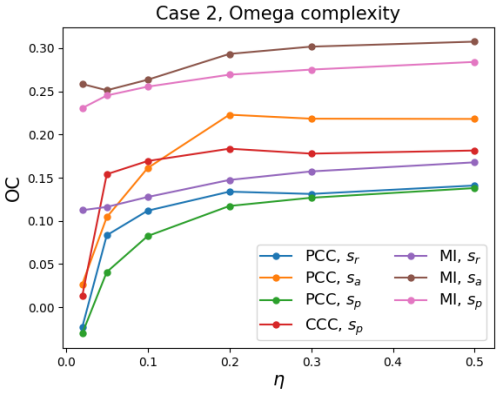
TE is used in two ways, both as a method to detect a change in information flow when there is a change in dependency in the system, but also as weight for a directed graph, which will be used for community detection.

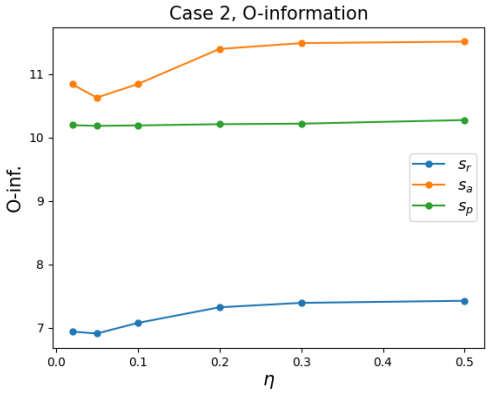
Coupled Rössler Systems

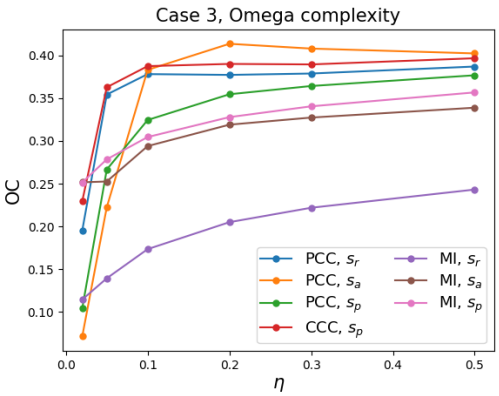
Since there are six cases and six different ηs for each case there are numerous results for the Rössler systems. The average over the TE results for all cases with $\eta = 0.1$ are shown in Figures 10.5-10.10 where the information flow from the i'th to the j'th signal is visualised. All results can be seen in Appendix B and the figures in this section are chosen as these most clearly show patterns in information flow.

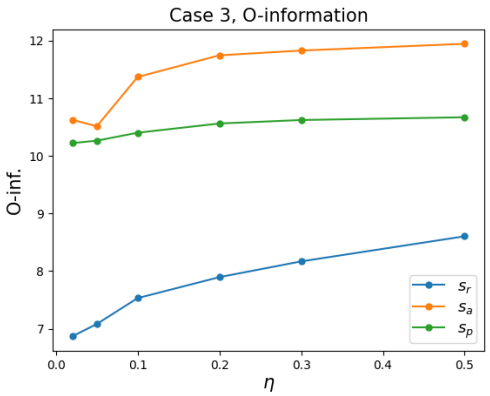

Case 1, Omega complexity
OC
η
PCC, s_r
PCC, s_a
PCC, s_p
CCC, s_p
MI, s_r
MI, s_a
MI, s_p

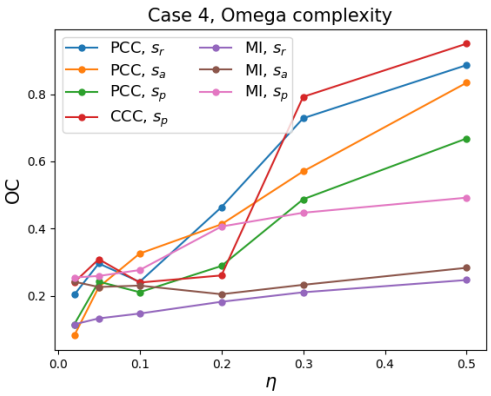

Case 1, O-information
O-inf.
η
s_r
s_a
s_p

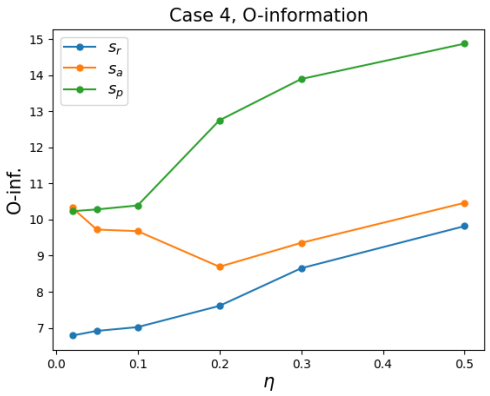

Case 2, Omega complexity
OC
η
PCC, s_r
PCC, s_a
PCC, s_p
CCC, s_p
MI, s_r
MI, s_a
MI, s_p


Case 2, O-information
O-inf.
η
s_r
s_a
s_p


Case 3, Omega complexity
OC
η
PCC, s_r
PCC, s_a
PCC, s_p
CCC, s_p
MI, s_r
MI, s_a
MI, s_p


Case 3, O-information
O-inf.
η
s_r
s_a
s_p


Case 4, Omega complexity
OC
η
PCC, s_r
PCC, s_a
PCC, s_p
CCC, s_p
MI, s_r
MI, s_a
MI, s_p


Case 4, O-information
O-inf.
η
s_r
s_a
s_p

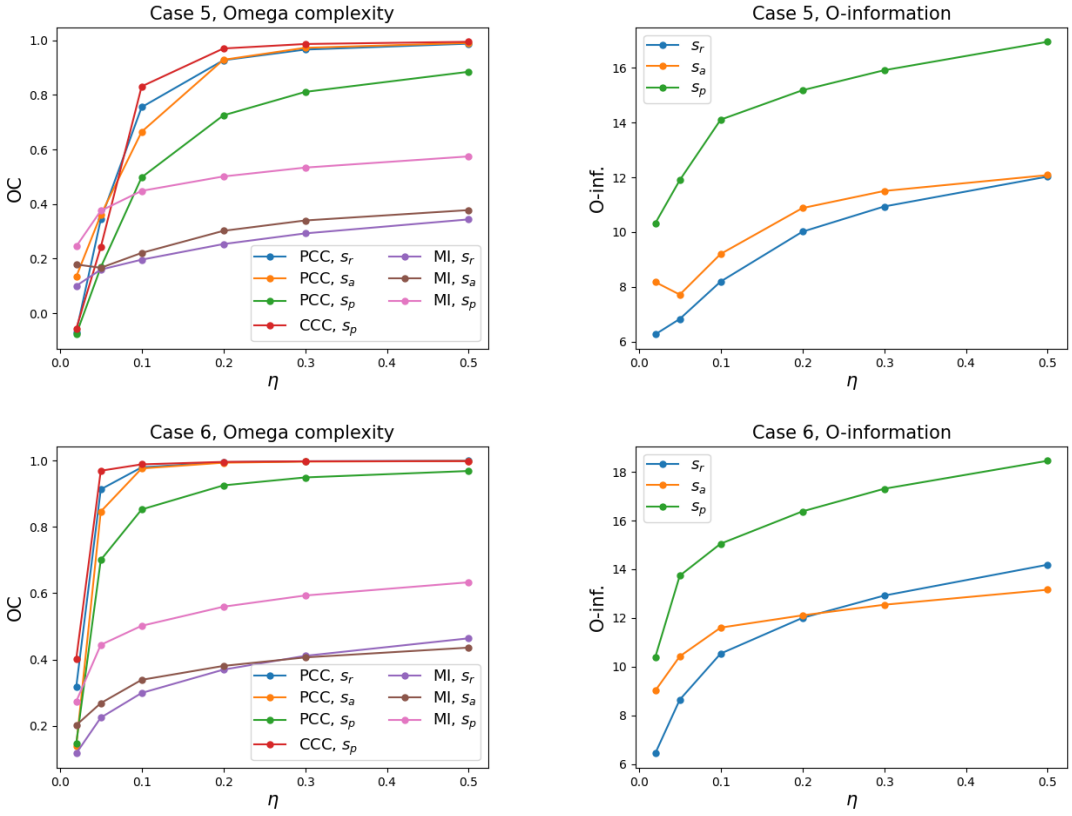


Figure 10.1: Means of OCs and O-information of Rössler systems coupled according to the six cases in Figure 9.1 with varying η.

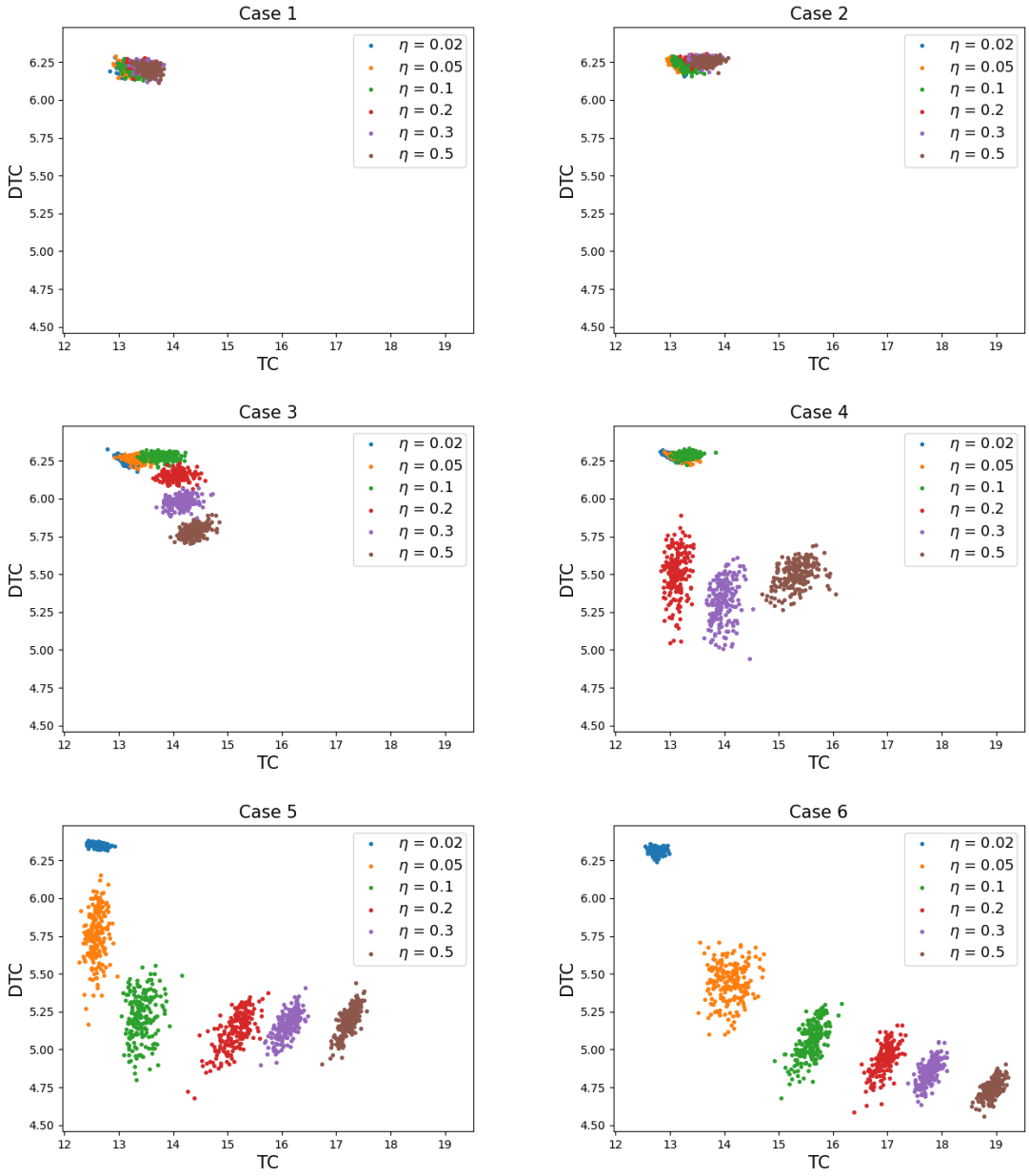


Figure 10.2: Total correlation and dual total correlation for the six cases of coupling of Rössler systems shown in Figure 9.1 for varying η.

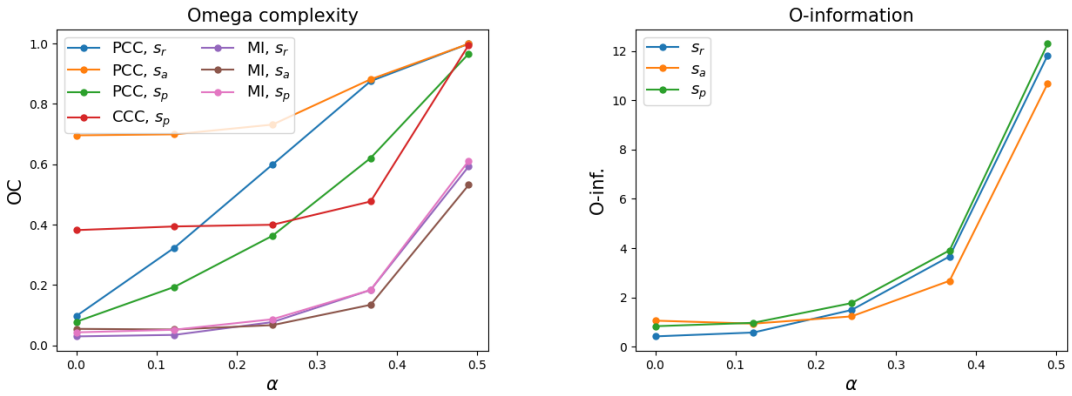


Figure 10.3: Means of OCs and O-information of MVAR processes as α increases.

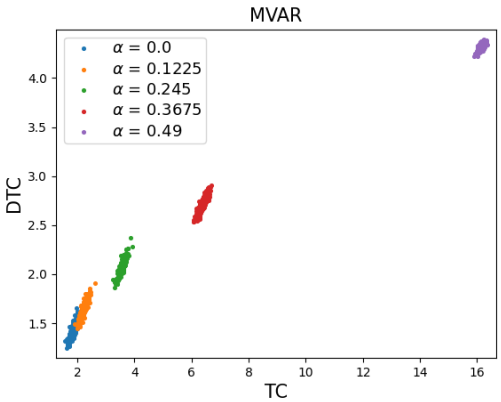


Figure 10.4: Total correlation and dual total correlation for the MVAR processes for varying α.

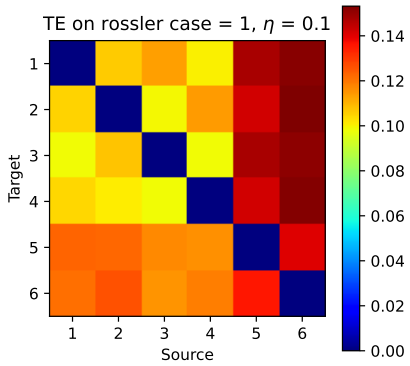


Figure 10.5: Transfer entropy on case 1 and $\eta = 0.1$

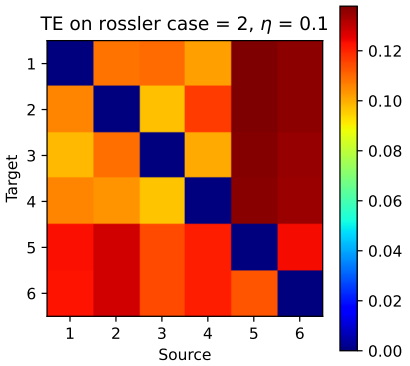


Figure 10.6: Transfer entropy on case 2 and $\eta = 0.1$

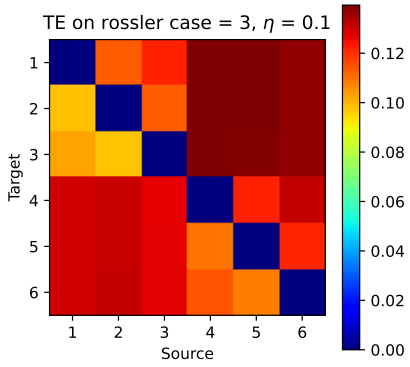


Figure 10.7: Transfer entropy on case 3 and $\eta = 0.1$

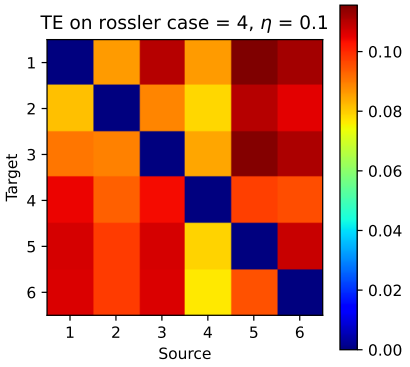


Figure 10.8: Transfer entropy on case 4 and $\eta = 0.1$

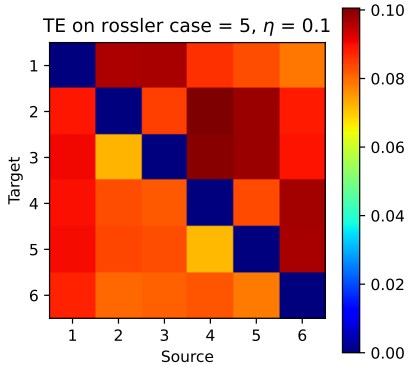


Figure 10.9: Transfer entropy on case 5 and $\eta = 0.1$

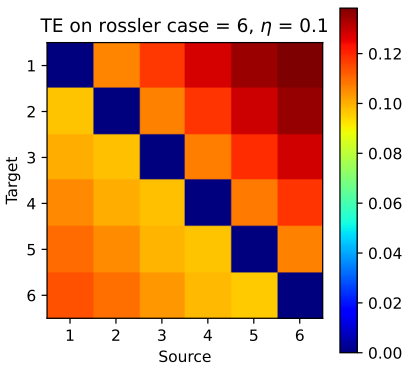


Figure 10.10: Transfer entropy on case 6 and $\eta = 0.1$

MVAR processes

The average over TE on all simulations of MVAR processes are presented in Figures 10.11-10.15.

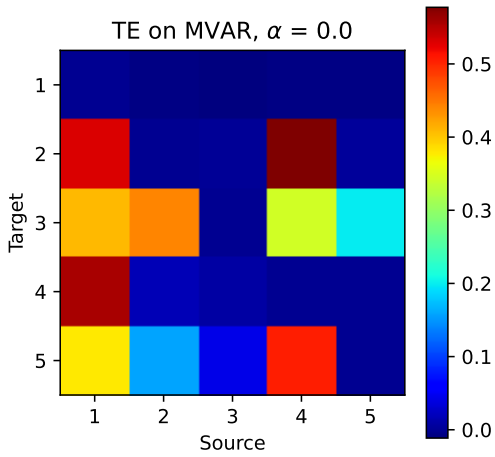


Figure 10.11: MVAR Process $\alpha = 0.0$

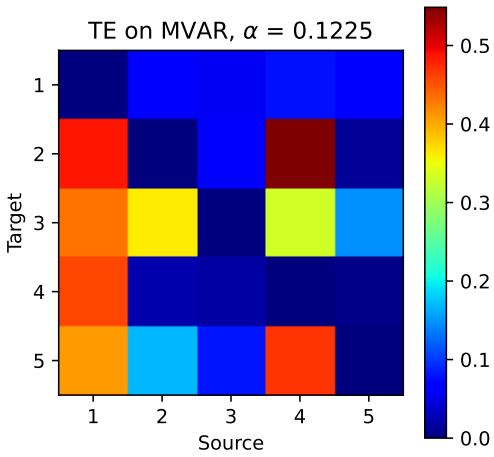


Figure 10.12: MVAR $\alpha = 0.1225$

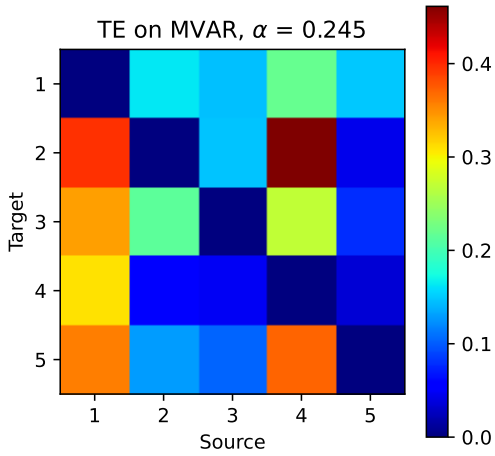


Figure 10.13: MVAR $\alpha = 0.245$

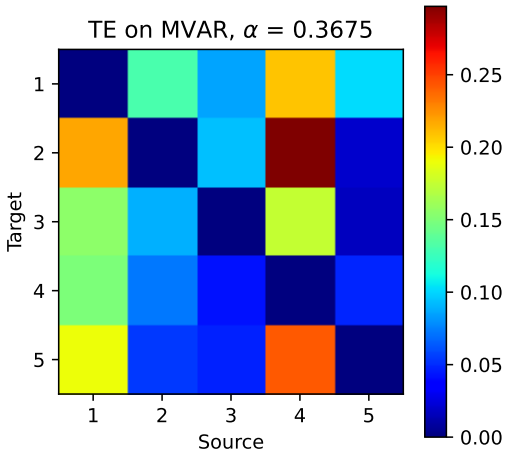


Figure 10.14: MVAR $\alpha = 0.3675$

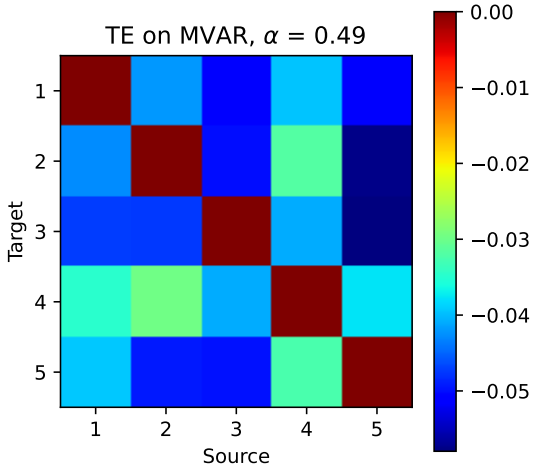


Figure 10.15: MVAR $\alpha = 0.49$

10.4.4 Directed Graph Using Transfer Entropy

Figure 10.16 shows the graphs constructed considering signals as nodes and the TEs as weights of directed edges. Darker color of edges indicates a higher transfer entropy and different colors of nodes indicate different clusters. Note that for $\alpha = 0.1225$ the Louvain algorithm is able to distinguish two communities 10.16, whereas $\alpha = 0.3675$ yields a single cluster. This is most likely due to the fact that all nodes are strongly connected after mixing and hence no clear cut communities are present.

This method could in theory be applied to the TE results from the coupled Rössler systems, but since no obvious ways of validating the output come to mind, this will not be explored.

The remaining results can be found in Appendix C.

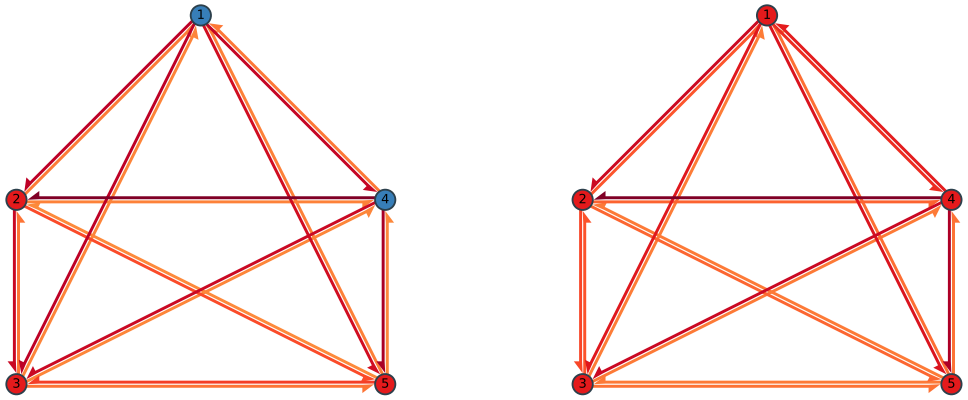


Figure 10.16: Clusterings of MVAR with $\alpha = 0.1225$ and $\alpha = 0.3675$.

10.5 Discussion of Simulated Results

The main takeaways from the results presented in Section 10.4 are summarised in this section.

10.5.1 Omega Complexities

Coupled Rössler Systems

In Figure 10.1 the leftmost plots show that the OC in general increases both as η increases and as the number of couplings increases regardless of dependency function. This is expected since each case of coupling increases the number of non-zero couplings and η determines the strength of these couplings. The only exception to this is for case 4, which shows decreasing $\Omega(\rho_p; s_r)$, $\Omega(\rho_p; s_p)$ and $\Omega(\rho_c; s_p)$ from $\eta = 0.05$ to $\eta = 0.1$. This could be a consequence of the coupling between nodes 2 and 5 which is the only coupling between the cluster of nodes 1, 2 and 3 and the cluster of nodes 4, 5 and 6 in Figure 9.1 causing irregular behaviour stemming from a weak ($\eta \in \{0.1, 0.2\}$) coupling between two clusters. The behaviour of the coupled Rössler systems as the number of couplings and the configuration of these change remains to be examined in more detail. Several systems with the same number of couplings but different configurations might show varying behaviour. This study will not be conducted here.

The OC based upon the MI in general has smaller range across cases than the OC based upon PCC and CCC. It is on average higher than the one based upon PCC and CCC for cases 1 and 2 and the inverse is true for cases 3, 4, 5 and 6. It in general fails to identify when signals are very similar. This can be seen from

Figure 9.2which shows that it changes the least from case 1 to case 6 where it is desirable that a large difference should be shown.

The OC based upon PCC of the instantaneous amplitude often has the largest range from $\eta = 0.01$ to $\eta = 0.5$ and furthermore most often monotonically increases in case number and η. Whether the increase in OC with respect to case number can be attributed to the configuration of coupling or the number of couplings is not clear. The characteristics of largest range and monotonic increase in case number and η are evaluated to be desirable when the goal is to measure changes in dependencies when the underlying characteristics of the system change.

Per definition the OC relies upon the dependency function and this must therefore be chosen according to assumptions regarding the processed signals. Since the OC based upon PCC and CCC reach values of almost 1, it seems that the dependencies between coupled Rössler systems can be explained by them, when the coupling is strong. Whether this is the case for low numbers and coefficients of coupling will not be examined here.

The OC is in general deemed able to discriminate between the six cases of coupling and and coupling degree $\eta \in \{0.01, 0.02, 0.1, 0.2, 0.3, 0.5\}$ of coupled Rössler systems from numerical solutions of them.

MVAR Processes

For the MVAR processes $\Omega(\rho_p; s_r)$ or $\Omega(\rho_p, s_p)$ produces the largest range of OC as α varies and it is furthermore monotonically increasing in α for the analysed points. The OC based upon MI again fails to identify the highly similar signals when $\alpha = 0.49$.

Notice that the graphs of the OC based upon MI are convex in the examined αs in Figure 10.3. The other graphs are not convex but close to. This is in contrast to the OCs of the Rössler systems for which most of them are concave in η with the notable exception of those for case 4. Note however that α and η do not express the same relationships for MVAR processes and Rössler systems, respectively, and the comparison is thus not direct even though they both represent some type of direct dependency.

10.5.2 TC, DTC and O-information

Coupled Rössler Systems

The O-information of the coupled Rössler systems in general increases in the six cases and in η with the exception of case 4 in which the O-information does not monotonically increase in the examined ηs shown in Figure 10.1. This is speculated to once again be due to the nature of the coupling configuration in case 4. Recall that the six cases have increasing number of couplings, and that it is speculated that this causes an increase in OC.

The O-information of s_r in general produces the largest range of values when η varies and furthermore most consistently increases in η, which suggests that this method in this case reflects the behaviour of the coupled Rössler systems better than the others examined.

As the O-information is positive in all cases shown and in general increases as η increases, it indicates that the coupled Rössler systems are dominantly redundant and that the redundancy of them increases in η.

In Figure 10.2 it is seen that increases in η become increasingly more influential on TC and DTC from case 1 to 6. This makes sense as the number of non-zero couplings increase and changing the coupling coefficient is therefore more influential. It is furthermore clear that TC in general increases in η and DTC decreases, indicating that the amount of synergy decreases and the redundancy increases.

MVAR Processes

The O-information of the MVAR processes is positive and increases in all examined αs as can be seen in Figure 10.3. The O-information therefore indicates that the MVAR processes are dominantly redundant and that the redundancy increases in α.

Figure 10.4 clearly shows that both TC and DTC increase in α but that TC increases much faster than DTC resulting in an increasingly positive O-information.

Notice that the shape of the plots of O-information in Figure 10.3 is different from the corresponding plots in Figure 10.1 showing a slightly different relationship between O-information and α compared to O-information and η for MVAR processes and coupled Rössler systems, respectively.

10.5.3 Transfer Entropy

Coupled Rössler systems

As can be seen from Figures 10.5-10.10 TE for the coupled Rössler systems shows very varying results. This could stem from the fact that the coupled Rössler systems are bi-directional as seen in Definition 9.2 and a symmetric coupling matrix ϵ is used, and therefore TE is an unfitting tool to apply to the Rössler systems.

MVAR processes

The TE captures information flow present in the MVAR process as seen in Figures 10.11-10.15, most clearly seen in Figure 10.11. Notice that a large information flow from nodes 4 to 2 is detected, and that this connection is not apparent in Equations 9.1. Signals 2 and 4 however appear to be strongly negatively correlated in Figure 9.3, thus the observed information flow from 4 to 2 could stem from this correlation. Additionally, in Figure 10.15, which is the case with highest mixture, most of the estimates are negative. This is obviously an error per Definition 5.16

which might stem from a bias or estimation error. This indicates, that the estimator of TE can yield counter intuitive results, when signals are very similar, as is the case with the MVAR process in Figure 10.15, where the high mixture of $\alpha = 0.49$ causes the individual signals to be almost identical.

11 | Analysis Considerations

In this chapter the considerations made before analysing EEG and intrusive EEG signals are presented. This includes introduction and discussion of choice of data sets, dependency measures, signal representations, analysis methods, significance tests, and clustering method.

11.1 Choice of Data Sets

Two different data sets, described in detail below, are analysed:

- SNR data set: A data set collected from 22 different test subjects performing two different tasks at two different levels of SNR.
- Seizure data set: A data set from a single epileptic subject with data obtained before (pre-ictal) and during (ictal) a seizure.

11.1.1 SNR Data Set

The first data set used was published alongside Hjortkjær et al. [2020] and is comprised of EEG measurements from 64 electrodes with a sampling frequency of 128 Hz, obtained from 22 healthy subjects aged 19-28. Each subject underwent a 1-back task and a 2-back task under an SNR level of 0 dB, denoted trial types 1 and 2, respectively. The subjects furthermore underwent the 1-back task and 2-back task under an SNR of 10 dB, denoted trial types 3 and 4, respectively. All four trial types were performed 10 times by each subject.

The following sections will compare results from trial type 1 with trial type 3 and results from trial type 2 with 4.

The data set is used due to the assumption that the same task performed under different SNR levels produce different EEG signals which could be detected through analysis of the EEG signals.

Pre-processing

As the SNR is not pre-processed, this is done before analysing it. This is done to ensure clean data without anomalies. Common anomalies in EEG signals can stem

from muscle movement, faulty electrodes or noise in signals. The pre-processing in this project is inspired by the pre-processing done by the authors of Hjortkjær et al. [2020] and is performed using the `EEGLAB` package in MATLAB.

Firstly, each of the 40 trials are truncated to the length of the shortest trial, and then all 64 signals are concatenated producing 64 signals in total for all 40 trials. This is done in order to reduce computation time, and the signals are post pre-processing split into separate signals. This affects the signals, as the concatenation results in discontinuities at the concatenation points. When the signals are separated after pre-processing, the last 256 samples of each signal are removed as they contained artefacts from the pre-processing. Any signals from faulty electrodes found are interpolated using spherical interpolation.

Afterwards, the signals are re-referenced to the average of all 64 signals. The signals are filtered with a notch filter from 48Hz to 52Hz to filter out the constant 50Hz noise which stems from the AC current present in all signals. Then a low-pass filter with cutoff frequency 1Hz and and high-pass filter with cutoff frequency 40Hz are applied. Subsequently independent component analysis (ICA) is used to decompose the signals, and the components are then visually inspected to detect anomalies. This is performed to remove components, which stem from muscle activity or noise instead of brain activity. From 64 signals and components, 6.7 signals and 5.5 components are on average removed from the data set. The data sets of subjects 1, 2, 4, and 18 are rejected for various reasons – either too much noise, inexplainable frequency anomalies or for requiring too much pre-processing.

Finally, 256 samples are removed in the beginning of each data set, as speech onset in the experiments was after two seconds of data recording.

11.1.2 Epileptic Seizure Data

The second data set analysed is the epileptic seizure data set published alongside Kramer et al. [2008]. The data set was recorded from a single subject at 400 Hz under pre-ictal and ictal states. The recordings were made as an intrusive EEG (iEEG) with 64 electrodes in a square grid and two in-depth electrode strips each with six electrodes resulting a total of 76 electrodes. Eight repeats of each state were recorded and each of these trials contains 10 seconds of data resulting in 4000 samples. The data was by the authors of Kramer et al. [2008] reviewed by a neurophysiologist, and no faulty electrodes were found. Before any analysis is done, every signal is band-pass filtered with cutoff frequencies of 1 and 50Hz as was done in Kramer et al. [2008]. In the following analysis the pre-ictal and ictal data sets will be sought discriminated.

The SNR data set is chosen for its relevance in relation to the problem statement of the project and the seizure data set may support or oppose findings from the SNR data set.

Use of other or additional data sets may provide additional insight into how the dependencies in EEG signals may change under different stimuli of the subjects.

11.2 Dependency Measures and Signal Representations

The EEG and iEEG signals are analysed according to the same combinations of dependency measures and signals representations as specified in Section 10.2.

11.2.1 Dependency Measures

The dependency measures OC and O-information are chosen for their different approaches to quantifying dependency in more than two variables as well as the feasibility of computing them in a fitting time frame. The OC and O-information have furthermore been used in literature with the OC being ubiquitous in analysis of EEG signals. The OC relies upon an aggregation of bivariate dependency functions while the O-information is non-bivariate.

The OC was analyzed in Chapter 7 and some advantages and disadvantages of it were documented. A generalised OC was introduced in Definition 7.10, which rectified some of these problems. For the reasons explained in Chapter 7 it was used instead of the original OC.

The choice of dependency function in the OC is important as it dictates which types of dependencies are registered as well as the difficulty of estimating the OC. The dependency functions used in the OC are chosen for the following reasons:

- PCC: It is simple, widely used, and computationally cheap.
- CCC: It is simple, computationally cheap and evaluates two signals which differ only in phase shifts to be more similar than the PCC does.
- MI: It captures more complex dependencies than linear coefficients such as PCC and CCC, while being computationally more complex.

Other dependency functions which could be used, include but are not limited to non-bivariate dependency functions allowing consideration of dependency between more than two variables. Examples include partial correlation, conditional mutual information, and interaction information.

O-information is by definition the difference between TC and DTC, which are both generalisations of MI. As shown in Section 5.4 O-information characterises the dependency in a set of variables as dominantly redundant or synergistic but provides no information about the magnitude of either type. A more detailed approach could be to utilise TC and DTC individually or introduce an O-information coefficient which relates the O-information to the magnitude of TC and DTC as the magnitude of the O-information is dependent on the variables treated.

All estimates of differential information theoretic quantities in the project are done based upon either the asymptotically unbiased kNN estimator of entropy or the KSG estimator of MI as documented in Chapter 6. The authors are aware of the bias of the methods, but its effects will not be considered further in this project.

11.2.2 Signal Representations

In addition to the real representation, the Hilbert transform, described in Chapter 2, is applied to all signals in order to extract instantaneous amplitudes and phases through the analytic representation of the signals. This is done to be able to assess the amplitude and phase information separately, inspired by previous work concerning phase synchrony and listening effort Baboukani et al. [2018]. It is furthermore motivated by the assumption that different representations of a signals facilitate easier extraction of relevant information.

Alternatively one could have examined other representations of the signals such as the frequency information through either the Fourier or wavelet transform, or envelopes constructed from alpha shapes.

11.3 Significance Tests

In order to test whether results from two different experiment setups are significantly different, they are compared using Welch's t-test first introduced in Welch [1947]. Definition 11.1 is based upon Sakai [2016].

Definition 11.1 (Welch's *t*-test)
Let X and Y be Gaussian stochastic variables with mean and variance (μ_X, σ_X) and (μ_Y, σ_Y), respectively. Let furthermore $\{x_k\}_{k=1}^{n_y}$ and $\{y_k\}_{k=1}^{n_y}$ be n_x and n_y i.i.d. samples of X and Y, respectively, and $\overline{x}$ and $\overline{y}$ their respective sample means. Welch's t-statistic is defined as

$$t_w = \frac{\overline{x} - \overline{y}}{\sqrt{v_x/n_x + v_y/n_y}},$$

where $v_x = \sum_{j=1}^{n}(x_j - \overline{x})^2/(n-1)$. The p-value is then

$$p = P(|T| \geq |t_w|)|\mu_X = \mu_Y),$$

where $T \sim t(\hat{\phi})$ and

$$\hat{\phi} = \left(\frac{v_x}{n_x} + \frac{v_y}{n_y}\right)^2 \Bigg/ \left(\frac{(v_x/n_x)^2}{n_x - 1} + \frac{(v_y/n_y)^2}{n_y - 1}\right).$$

In this setup the null hypothesis is that the means of the results from two different experimental setups are equal.

Welch's t-test is chosen for various reasons:

- It is relatively simple.

- It is assumed that the results are approximately normally distributed.
- It is assumed that the distribution of two sets of results being compared have unequal variances or that it is at least not reasonable to make an assumption of equal variances.

Instead of merely comparing means, the variances could be estimated, such that likelihoods and posteriors could be used to characterise differences in results.

Since multiple t-tests are made, the results are subject to the problem of false discovery rate in multiple comparisons [Benjamini and Hochberg, 1995].

Definition 11.2 (False Discovery Rate (FDR))
Consider simultaneous testing of multiple null-hypotheses of which some are true. If V is the number of wrongly rejected null-hypotheses and R is the total number of rejected null-hypotheses the false discovery rate is $E[V/R]$.

It is desirable to lower the FDR in order to better reflect which null-hypotheses should be rejected. In two cases treated in this project, multiple p-values are obtained through multiple comparisons:

- When Welch's t-test is applied to results obtained from multiple subjects, the null-hypothesis is repeated for each subject.
- When Welch's t-test is applied to results obtained from multiple clusters of the same data set, the null-hypothesis is repeated for each cluster.

The Benjamini-Hochberg method is in this project used to bound the FDR [Benjamini and Hochberg, 1995].

Definition 11.3 (Benjamini-Hochberg Method)
Consider testing m null-hypotheses H_i with corresponding p-values P_i and the ordering of the p-values $P_{(1)} \leq \ldots \leq P_{(m)}$ with corresponding null-hypotheses $H_{(i)}$. The Benjamini-Hochberg method with control rate $q \in [0, 1]$ is defined as the rejection of all $H_{(i)}$, $i = 1, \ldots, k$, where

$$k = \underset{j}{\operatorname{argmax}} \left(P_{(j)} \leq \frac{j}{m} q \right).$$

The Benjamini-Hochberg method is in Benjamini and Hochberg [1995] shown to bound the FDR by q, and in this project $q = 0.05$ is used.

The Benjamini-Hochberg method is used for its ease of implementation and upper bound of FDR at q. The choice of $q = 0.05$ is from convention, and the method is implemented with the `statsmodels` package in Python. Other methods have not been considered, but could be chosen depending on the desired control of the errors of the hypothesis testing.

11.4 Clustering Method

The data sets described in Sections 11.1.1 and 11.1.2 are first analysed by considering all signals from all electrodes – that is 64 for the SNR data set and 76 for the seizure data set. Subsequently, two different clusterings of electrodes in the SNR data set and three different clusterings of electrodes in the seizure data set are analysed.

These clusterings are constructed with the Louvain clustering algorithm as described in Section 8.3.1. The SNR data set is clustered with dependency matrices from PCC and MI, while the seizure data set is clustered with dependency matrices from PCC, MI and TE. TE was not used as a clustering method for the SNR data set as the results varied to a degree that made them unfit for further analysis and hence not suitable for clustering. The bivariate dependency functions PCC, MI and TE used in conjunction with Louvain for clustering were chosen as a consequence of their different ways of quantifying dependencies and due to time constraints. Notice that the definition of TE requires stationarity of the signals treated. It is assumed that the EEG and iEEG signals are stationary in small time periods which justifies the use of TE with fitting time lags.

The clusterings are constructed with comparisons in mind. Thus a combined clustering is produced for every pair of data type which are to be compared. This is done according to the steps in Section 8.4 and the data sets are then analysed according to their combined clustering.

The clusterings are denoted c_{PCC}, c_{MI} and c_{TE} when constructed with PCC, MI and TE, respectively, and the data set which is clustered is specified from context. The individual clusters from each clustering method are referred to by the dependency matrix upon which they are based and a color with which the specific cluster is identified – for example $c_{PCC,blue}$, which specifies the blue cluster from the clustering made with the Pearson correlation coefficient. Clusterings obtained from the data set of a specific subject are never used for other subjects than the one it was constructed from.

The clusterings from PCC and MI are only analysed with OCs $\Omega(\rho_p; s_r)$ and $\Omega(I; s_r)$, respectively, so as to retain consistency of the methods. The O-information $\mathcal{O}(s_r)$ is applied on all clusters. In contrast, the clusterings from TE are analysed according to all combinations in Section 10.2 as TE has no direct equivalent in the list of dependency measures and signal representations.

The Louvain clustering algorithm is chosen because of its generality in terms of directed and undirected graphs as well as its ability to adjust the number of clusters to increase the modularity of the clustering. Different bivariante dependency functions will likely result in clusterings which differ not only in the cluster configurations but also the number of clusters. The connection between the characteristics of the bivariate dependency functions and the resulting clusterings in relation to these will not be covered in detail here.

Two main considerations are made concerning the selection of nodes to analyse based on the clusterings:

1. Clusters of electrodes being compared must include the same nodes. This is done in order to only consider nodes which are relevant in both representations of the same community.

2. Clusterings of two different data sets which overlap at a large number of nodes are deemed representative of the same underlying community.

It is assumed that the above method of clustering will facilitate comparisons of electrodes which better capture specific differences when a subject undergoes different conditions.

This project only considers clusters obtained from the Louvain algorithm, but other clustering methods may uncover different clusters if applied to the dependency matrices.

11.5 Analysis Method

The results are obtained through mostly data driven methods. The following methods are not data driven:

- The pre-processing of the SNR data set which relies upon prior knowledge of the nature of EEG signals and of which unwanted artefacts and noise types can appear in such data sets.

- The filtering of the seizure data set, which is done in alignment with the source with which the data was published.

- The discarding of samples in the beginning and ending of the recorded signals which is motivated by the fact that speech onset in the SNR experiment was two seconds after the start of the experiment and that artifacts from the pre-processing are present in both ends.

The results from all dependency measures, the significance tests and the clusterings will all be obtained independent of prior neurophysiological knowledge. This might be a disadvantage in the sense that important insights from neurophysiology are missed and some parts, for example the clusterings, might be made easier with prior knowledge.

12 | Results from EEGs and iEEGs

In this chapter results from analysis of EEG and iEEG signals are presented. The analysis is done as specified in Chapter 11. These results will in chapter 13 be used to evaluate whether significant differences between EGG signals obtained under different conditions can be found by use of the dependency measures presented in Chapters 2, 3, 5 and 7 in conjunction with different signals representations.

12.1 SNR Data Set Results

The OCs and the O-information of all trials for subject 3 are seen in Figure 12.1 and Figure 12.2, respectively. The patterns and main conclusions drawn from these results are similar across all subjects and the plots of the remaining results can be seen in Appendix D.

A black dashed line of gradient 1 with intersection in $(0, 0)$ is shown in the figures. If a data point is located on the line, it indicates equal values for the specific duplicate of both trial types. If a point is located above, it indicates a larger value for the trial type on the vertical axis and vice versa if the point is below. Thus, if the mean of the values is centered near this diagonal line, it indicates that they are not different.

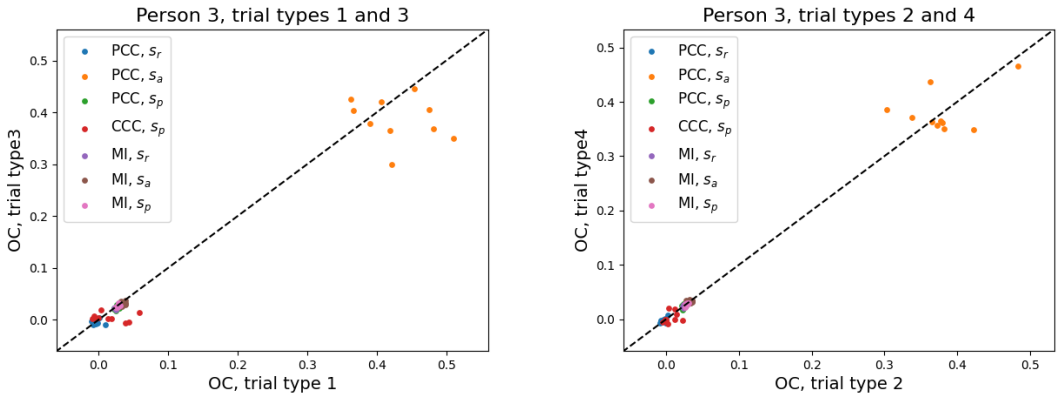


Figure 12.1: OCs of all duplicates of trial types 1 and 3 and trial types 2 and 4, respectively, for subject 3.

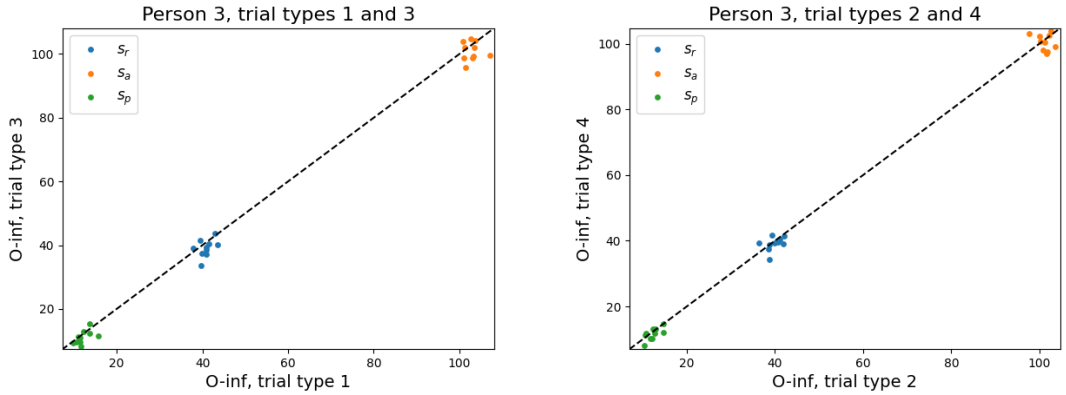


Figure 12.2: O-information of all duplicates of trial types 1 and 3 and trial types 2 and 4, respectively, for subject 3.

Welch's t-test with Benjamini-Hochberg correction of the p-values across all 18 subjects for the comparison of trial types 1 and 3 and trial types 2 and 4 is performed. None of the results across all subjects and all tools used for quantifying dependencies were significant at a significance level of 5% after correction.

12.1.1 Clustered SNR Data Set

The combined clusterings c_{PCC} and c_{MI} of trial types 1 and 3 based upon PCC and MI for both subject 3 and 7 are seen in Figures 12.3-12.6.

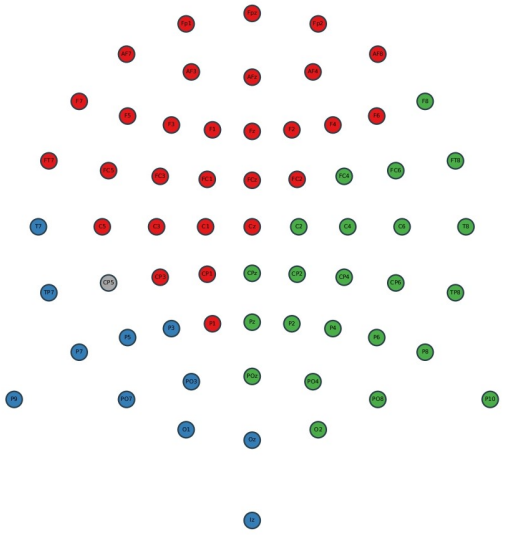


Figure 12.3: Combined clustering c_{PCC} for subject 3 and trial types 1 and 3.

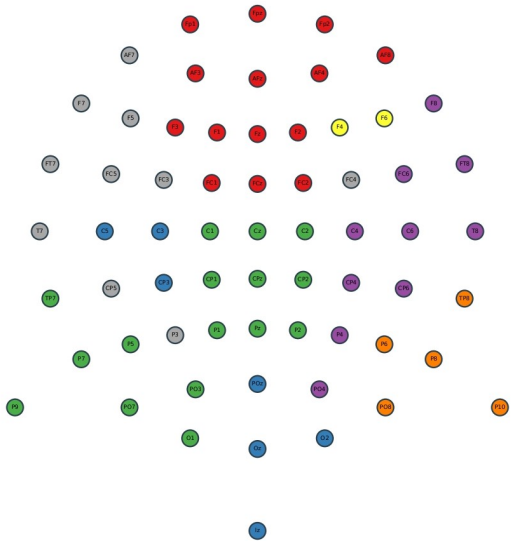


Figure 12.4: Combined clustering c_{MI} for subject 3 and trial types 1 and 3.

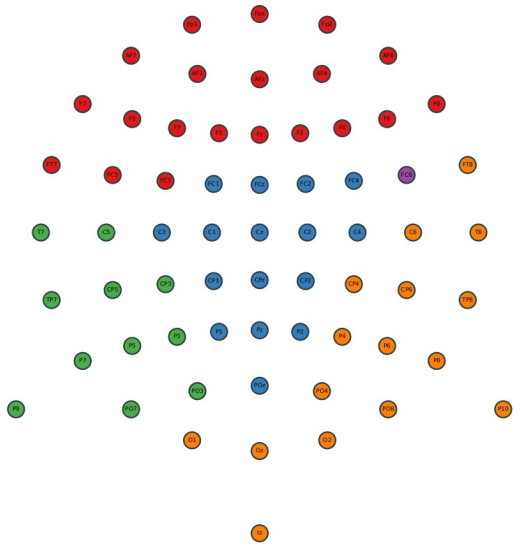


Figure 12.5: Combined clustering c_{PCC} for subject 7 and trial types 1 and 3 constructed with PCC.

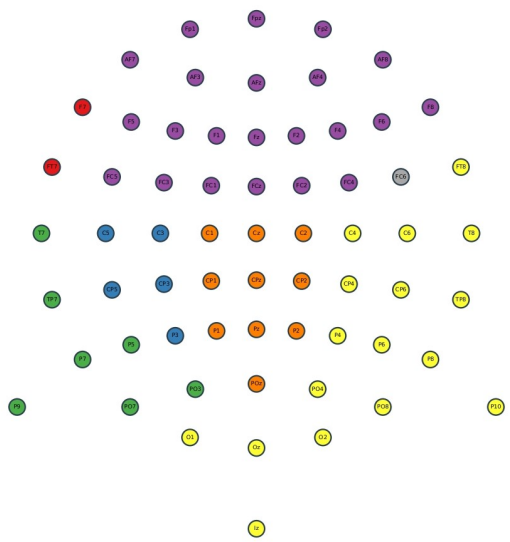


Figure 12.6: Combined clustering c_{MI} for subject 7 and trial types 1 and 3 constructed with MI.

The results from applying $\Omega(\rho_p; s_r)$ and $\Omega(I; s_r)$ on the clusterings c_{PCC} and c_{MI}, respectively, on the data set from trial types 1 and 3 from subject 3 are seen in Figures 12.7-12.10. Corresponding results for subject 7 are seen in Figures 12.11-12.14 with the clusterings c_{PCC} an c_{MI} based upon the data sets from trial types 1 and 3.

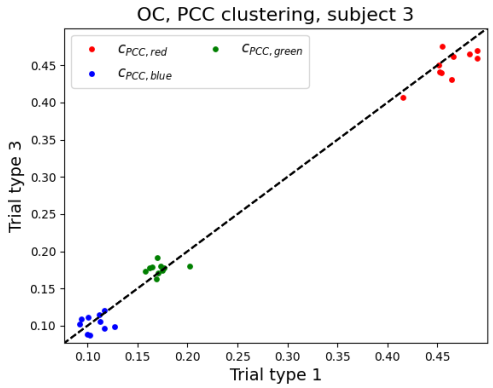


Figure 12.7: OCs $\Omega(\rho_p; s_r)$ of the clustering c_{PCC} for subject 3.

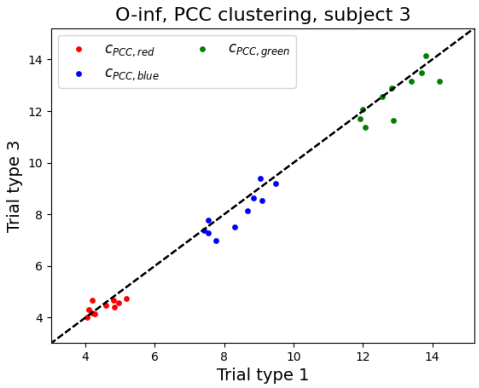


Figure 12.8: O-information $\mathcal{O}(s_r)$ of the clustering c_{PCC} for subject 3.

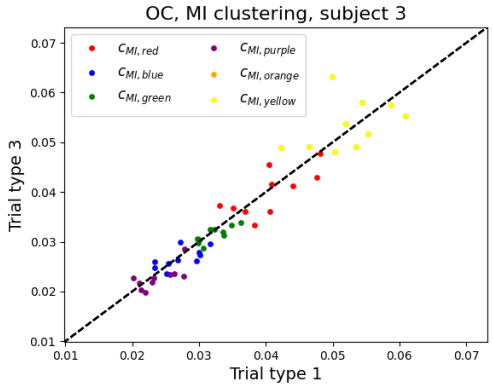


Figure 12.9: OCs $\Omega(I; s_r)$ of the clustering c_{MI} for subject 3.

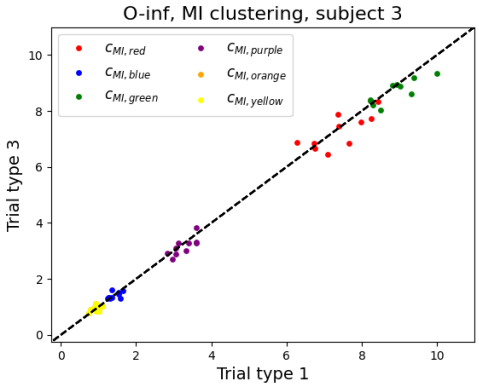


Figure 12.10: O-information $\mathcal{O}(s_r)$ of the clustering c_{MI} for subject 3.

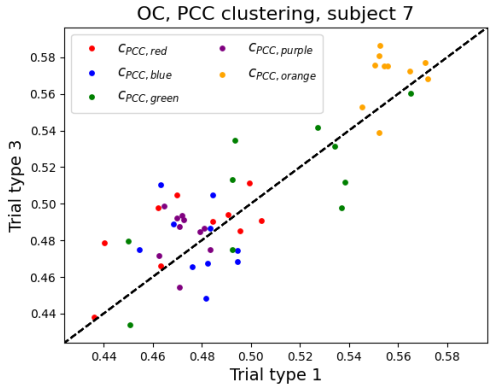


Figure 12.11: OCs $\Omega(\rho_p; s_r)$ of the clustering c_{PCC} for subject 7.

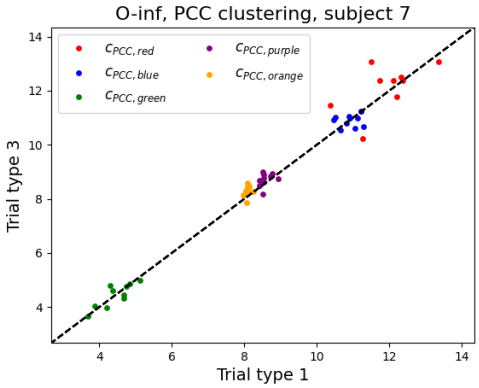


Figure 12.12: O-information $\mathcal{O}(s_r)$ of the clustering c_{PCC} for subject 7.

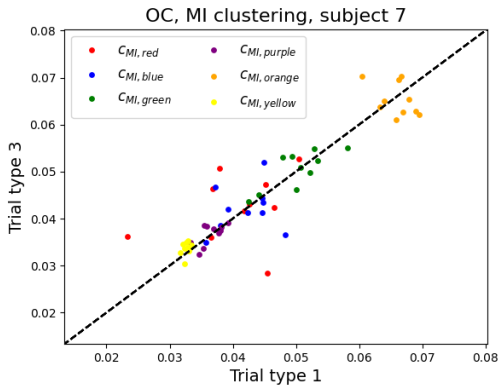


Figure 12.13: OCs $\Omega(I; s_r)$ of the clustering c_{MI} for subject 7.

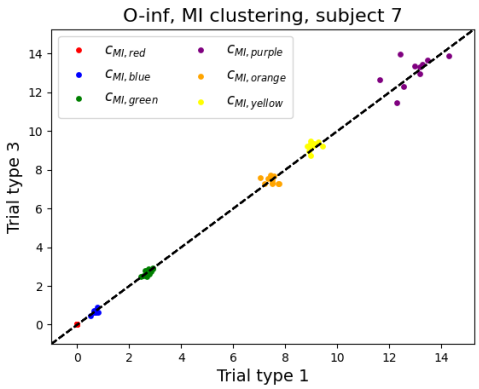


Figure 12.14: O-information $\mathcal{O}(s_r)$ of the clustering c_{MI} for subject 7.

Welch's t-test with Benjamini-Hochberg correction across clusters yielded no significant differences between trials types 1 and 3 at a 5% significance level.

12.2 Seizure Data Set Results

The results from the seizure data set are seen in Figure 12.15 and 12.16.

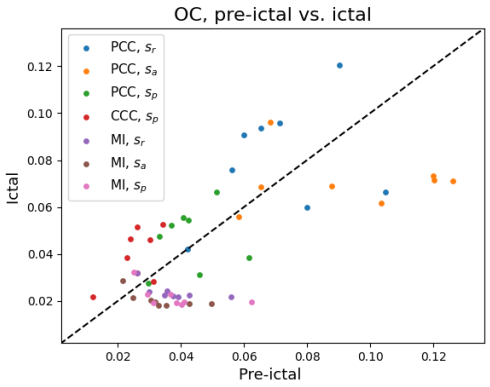


Figure 12.15: Comparison of the OC of the eight duplicates of the pre-ictal and ictal data sets.

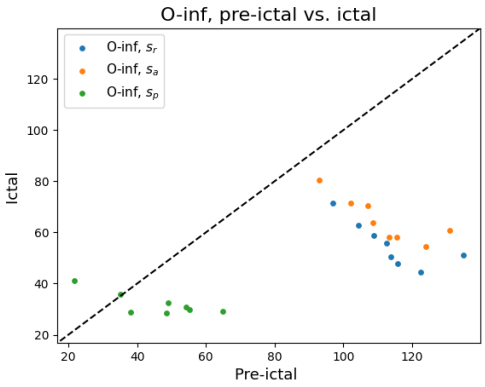


Figure 12.16: Comparison of the O-information of the eight duplicates of the pre-ictal and ictal data sets.

The results of Welch's t-test used for comparing the means of the results from the same dependency measure on pre-ictal and ictal data sets can be seen in Table 12.1.

Dep. measure	p-value	Sig. (5%)
$\Omega(\rho_p; s_r)$	4.15e-01	
$\Omega(\rho_p; s_a)$	5.73e-02	
$\Omega(\rho_p; s_p)$	5.21e-01	
$\Omega(\rho_c; s_p)$	1.34e-02	✓
$\Omega(I; s_r)$	2.83e-03	✓
$\Omega(I; s_a)$	4.09e-03	✓
$\Omega(I; s_p)$	4.08e-03	✓
$\mathcal{O}(s_r)$	3.49e-08	✓
$\mathcal{O}(s_a)$	7.07e-07	✓
$\mathcal{O}(s_p)$	2.43e-02	✓

Table 12.1: Dependency measures of the seizure dataset and their corresponding p-values from Welch's t-test comparing pre-ictal and ictal data sets.

12.2.1 Clustered Seizure Data Set

The combined clusterings c_{PCC}, c_{MI} and c_{TE} on the seizure data set are seen in Figures 12.17-12.19. The results from the seizure data set clustered with PCC and MI are seen in Figures 12.20-12.23. The results from the seizure data set clustered with TE are seen in Figures 12.24-12.33.

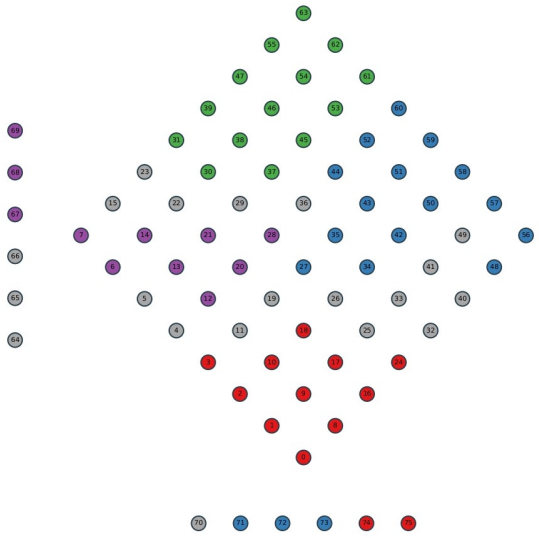


Figure 12.17: Combined clustering c_{PCC}. Grey nodes are excluded nodes, i.e nodes which do not belong to any combined cluster.

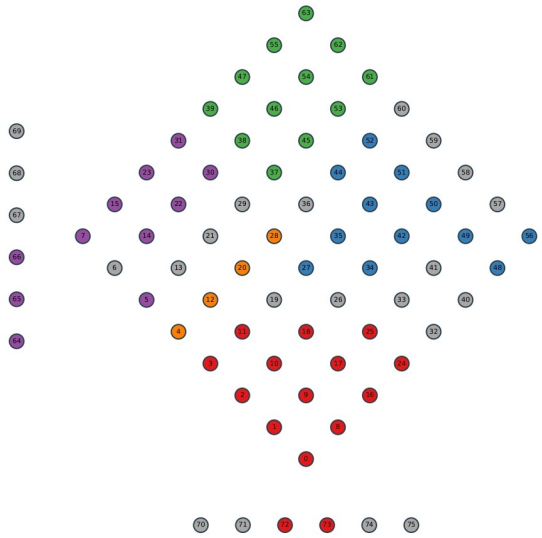


Figure 12.18: Combined clustering c_{MI}. Grey nodes are excluded nodes, i.e nodes which do not belong to any combined cluster.

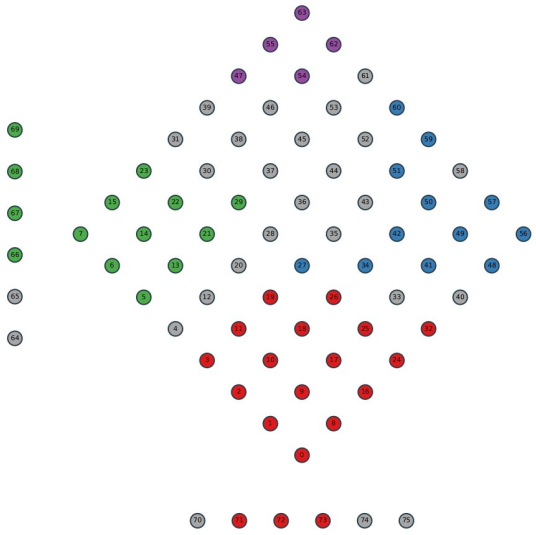


Figure 12.19: Combined clustering c_{TE}. Grey nodes are excluded nodes, i.e nodes which do not belong to any combined cluster.

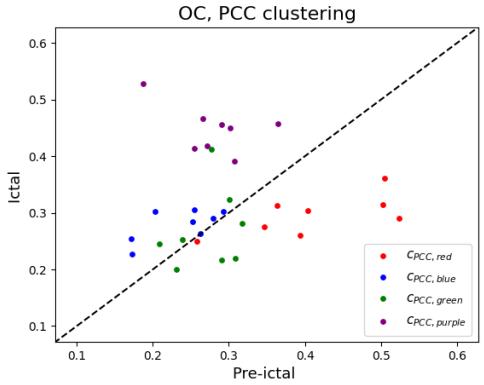


Figure 12.20: Omega complexities $\Omega(\rho_p; s_r)$ of the clusters based upon the PCC matrix of s_r.

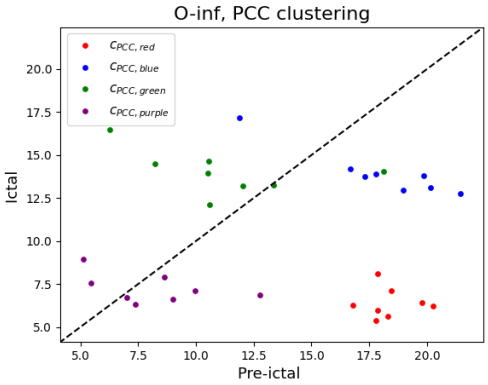


Figure 12.21: O-information $\mathcal{O}(s_r)$ of the clusters based upon the PCC matrix of s_r.

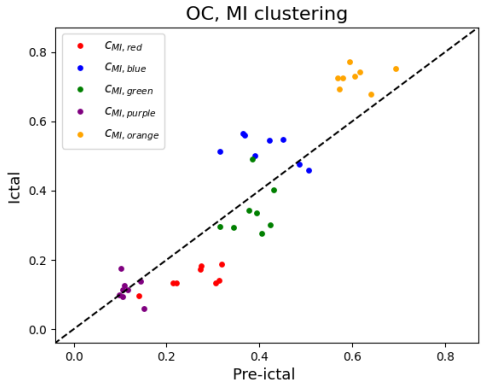


Figure 12.22: Omega complexities $\Omega(I; s_r)$ of the clusters based upon the MI correlation matrix of s_r.

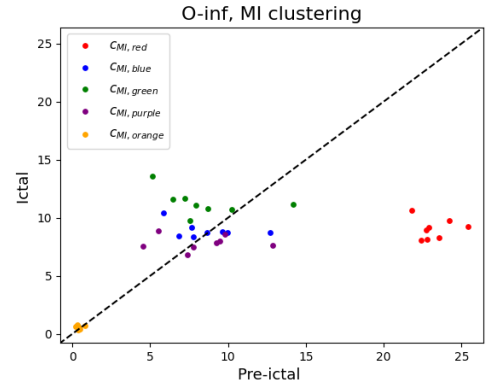


Figure 12.23: O-information $\mathcal{O}(s_r)$ of the clusters based upon the PCC matrix of s_r.

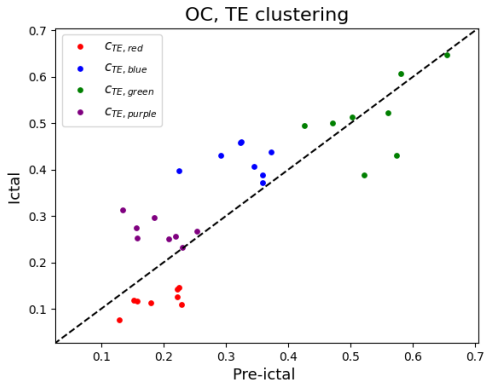


Figure 12.24: Omega complexity $\Omega(\rho_p; s_r)$ of the clusters in c_{TE}.

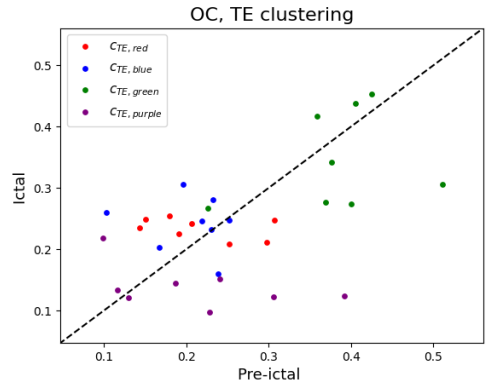


Figure 12.25: Omega complexity $\Omega(\rho_p; s_a)$ of the clusters in c_{TE}.

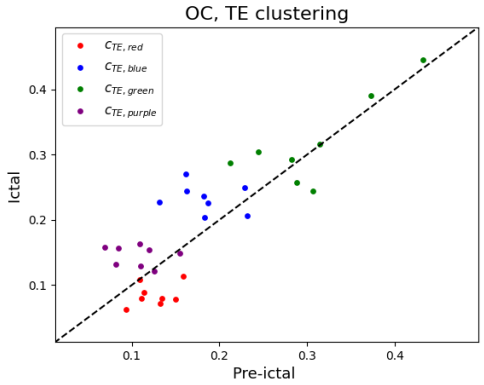


Figure 12.26: Omega complexity $\Omega(\rho_p; s_p)$ of the clusters in c_{TE}.

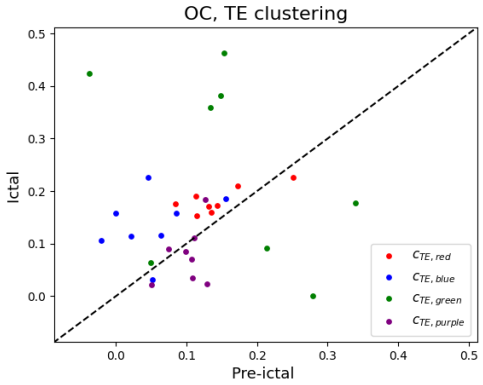


Figure 12.27: Omega complexity $\Omega(\rho_c; s_p)$ of the clusters in c_{TE}.

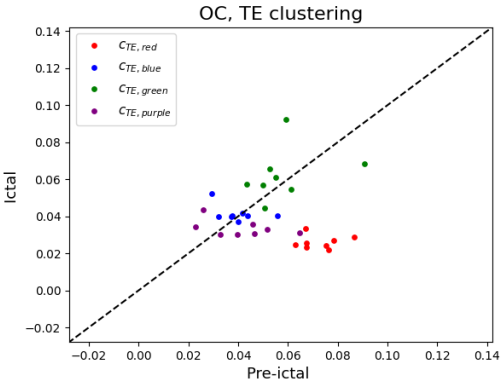


Figure 12.28: Omega complexity $\Omega(I; s_r)$ of the clusters in c_{TE}.

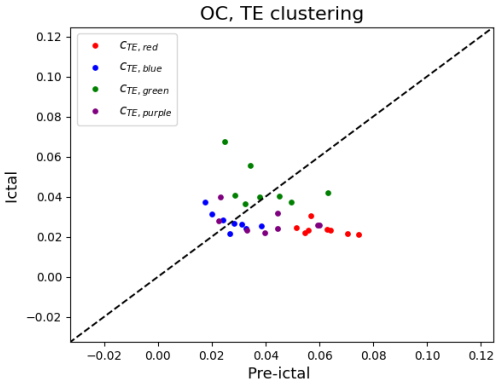


Figure 12.29: Omega complexity $\Omega(I; s_a)$ of the clusters in c_{TE}.

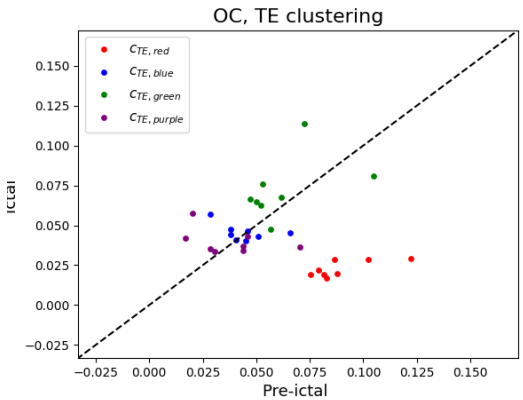


Figure 12.30: Omega complexity $\Omega(I; s_p)$ of the clusters in c_{TE}.

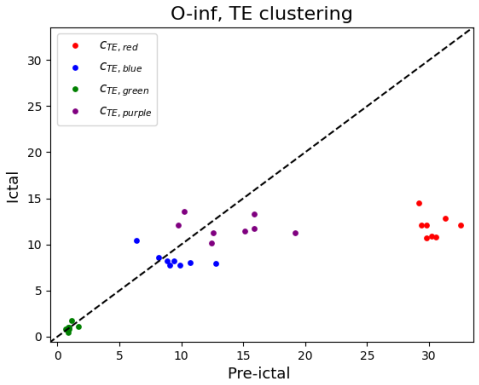


Figure 12.31: O-information $\mathcal{O}(s_r)$ of the clusters in c_{TE}.

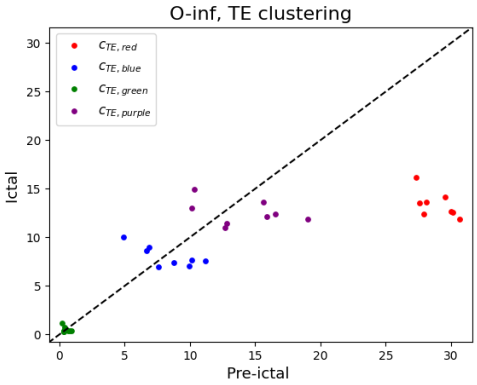


Figure 12.32: O-information $\mathcal{O}(s_a)$ of the clusters in c_{TE}.

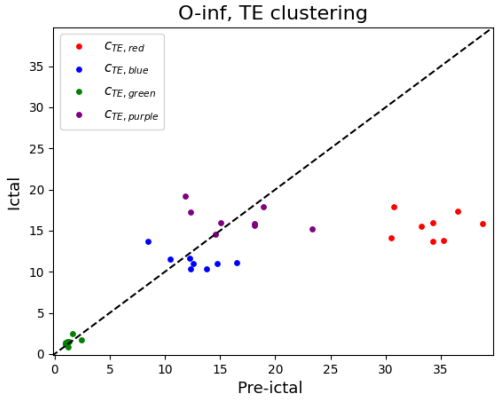


Figure 12.33: O-information $\mathcal{O}(s_p)$ of the clusters in c_{TE}.

The dependency measures, clustering methods, and corresponding p-values obtained through Welch's t-test corrected through the Benjamini-Hochberg method are seen in Tables 12.2, 12.3 and 12.4 for c_{PCC}, c_{MI} and c_{TE}, respectively.

Dep. measure	Cluster	p-value	Sig. (5%)
$\Omega(\rho_p; s_r)$	$c_{PCC,red}$	1.80e-02	✓
$\Omega(\rho_p; s_r)$	$c_{PCC,blue}$	6.50e-02	
$\Omega(\rho_p; s_r)$	$c_{PCC,green}$	9.23e-01	
$\Omega(\rho_p; s_r)$	$c_{PCC,purple}$	2.18e-05	✓
$\mathcal{O}(s_r)$	$c_{PCC,red}$	1.45e-11	✓
$\mathcal{O}(s_r)$	$c_{PCC,blue}$	1.08e-02	✓
$\mathcal{O}(s_r)$	$c_{PCC,green}$	8.59e-02	
$\mathcal{O}(s_r)$	$c_{PCC,purple}$	3.54e-01	

Table 12.2: Dependency measures, clusters in c_{PCC} and their corresponding p-values from Welch's t-test comparing pre-ictal and ictal data sets.

Dep. measure	Cluster	p-value	Sig. (5%)
$\Omega(I; s_r)$	$c_{MI,red}$	2.58e-03	✓
$\Omega(I; s_r)$	$c_{MI,blue}$	3.27e-03	✓
$\Omega(I; s_r)$	$c_{MI,green}$	2.19e-01	
$\Omega(I; s_r)$	$c_{MI,purple}$	9.28e-01	
$\Omega(I; s_r)$	$c_{MI,yellow}$	1.11e-04	✓
$\mathcal{O}(s_r)$	$c_{MI,red}$	2.06e-12	✓
$\mathcal{O}(s_r)$	$c_{MI,blue}$	7.32e-01	
$\mathcal{O}(s_r)$	$c_{MI,green}$	5.99e-02	
$\mathcal{O}(s_r)$	$c_{MI,purple}$	7.32e-01	
$\mathcal{O}(s_r)$	$c_{MI,yellow}$	6.99e-02	

Table 12.3: Dependency measures, clusters in c_{MI} and their corresponding p-values from Welch's t-test comparing pre-ictal and ictal data sets.

Dep. measure	$c_{TE,red}$		$c_{TE,blue}$		$c_{TE,purple}$		$c_{TE,green}$	
	p-value	Sig (5%)	p-value	Sig (5%)	p-value	Sig (5%)	p-value	Sig (5%)
$\Omega(\rho_p; s_r)$	1.52e-03	✓	1.52e-03	✓	5.62e-01		1.52e-03	✓
$\Omega(\rho_p; s_a)$	4.65e-01		2.76e-01		4.65e-01		2.76e-01	
$\Omega(\rho_p; s_p)$	5.73e-03	✓	7.57e-03	✓	7.61e-01		7.57e-03	✓
$\Omega(\rho_c; s_p)$	1.56e-01		3.46e-02	✓	2.93e-01		2.93e-01	
$\Omega(I; s_r)$	1.70e-07	✓	5.88e-01		5.88e-01		3.48e-01	
$\Omega(I; s_a)$	3.16e-06	✓	9.34e-01		4.86e-01		7.41e-02	
$\Omega(I; s_p)$	6.23e-06	✓	7.44e-01		6.02e-01		7.44e-01	
$\mathcal{O}(s_r)$	2.15e-13	✓	1.42e-03	✓	9.69e-01		2.49e-01	
$\mathcal{O}(s_a)$	5.19e-12	✓	9.68e-01		9.68e-01		4.28e-01	
$\mathcal{O}(s_p)$	1.42e-08	✓	4.23e-01		7.67e-01		9.46e-01	

Table 12.4: Dependency measures, clusters in c_{TE} and their corresponding p-values from Welch's t-test comparing pre-ictal and ictal data sets.

13 | Discussion

In this chapter the results presented in Chapter 12 are discussed.

All pairs of results which were compared through Welch's t-test were assumed to be approximately normally distributed. From visual inspection of the figures in Chapter 12 this can not be confirmed nor rejected, but it seems reasonable to believe that the distributions of the points in the scatter plots can be modelled with a normal distribution on each axis.

13.1 SNR Data Set

The SNR data set was analysed to see if different levels of SNR for equal experiment setups would be detectable with the different dependency measures and signal representations from Section 10.2.

13.1.1 All Electrodes

The results from the analysis of signals from all electrodes of the SNR data set provided by the dependency measures and signal representations from Section 10.2 resulted in no significant differences in means at a 5% significance level between trial types 1 and 3 and trial types 2 and 4, respectively.

Multiple reasons as to why this is the case are possible:

- There is simply no significant change in activity in the brains of the subjects from one trial type to another detectable through EEG signals. This is not believed to be the case.
- Considering 64 electrodes with the omega complexity means a total of $\sum_{i=1}^{64} i = 2080$ bivariate dependencies. Thus any noteworthy changes of a small subset of electrodes between two trial types might vanish in the overwhelming amount of dependencies which do not change much. If this is the case, analysis of clusters of electrodes might be a solution.
- In extension to the previous point, the omega complexities in Figure 12.1 are all with the exception of $\Omega(\rho_p; s_a)$ very low. It is assumed that most of the 2080 dependencies are very small since not all time series are assumed to be similar.

As covered in Chapter 7 small entries in dependency matrices will in general result in small omega complexities and a few high bivariate dependencies will not cause a noteworthy change in omega complexity.

- The dependency measures and signal representations are unsuited for detecting an eventual change in the EEG signals from one trial type to another.
- The pre-processing of the data set in some way decreased any significant change from one trial type to another.
- The problem of volume conduction mentioned in Chapter 1 may affect the dependency analysis of the EEG signals. The specific effect of this on the dependency analysis is unknown but might worsen the results.

The O-information is in all cases positive indicating that the data set is dominantly redundant.

13.1.2 Clustered Electrodes

The clusterings c_{PCC} and c_{MI} of trial types 1 and 3 for subjects 3 and 7 in the SNR data set seen in Figures 12.3-12.6 resulted in a different number of clusters for each bivariate dependency function used. The clusterings c_{PCC} and c_{MI} were analysed with $\mathcal{O}(s_r)$ and with $\Omega(\rho_p; s_r)$ and $\Omega(I; s_r)$, respectively.

The results from these clusterings produced no results with significant differences in means between trials 1 and 3 at a 5% significance level. Multiple reasons as to why this is the case are possible – the same points as touched upon in Section 13.1.1 apply as well as a few additional:

- The assumption that clusters of electrodes might provide more insight into differences between trial types 1 and 3 might be erroneous.
- The clusterings are purely data driven and might not correspond to any meaningful clustering in a neurophysiological sense. Parallels between these clusterings and those made in neurophysiology might drawn.
- The method of choosing shared clusterings might cause survivorship bias. That is, the assumption that the discriminating information is contained within the overlap between clusters is incorrect, and the relevant information is instead contained exactly in the excluded nodes.

Different results might have been obtained if a different clustering method had been used.

13.1.3 Pre-processing

As stated before, the SNR dataset was not pre-processed which meant that the authors of the project had to carry out the pre-processing. This was done with

inspiration from the pre-processing carried out in Hjortkjær et al. [2020] and to the best of the knowledge of the authors, but it should be stated that none of the authors had any prior experience with processing of EEG data.

As a result it can not be denied that the pre-processing has had any influence on the results and in extension the significance of the results. Hence based on the current experiment the authors do not feel confident rejecting or accepting the null hypotheses specified in Section 11.3 that there is no significant difference in OCs and O-information between trial 1 and trial 3.

13.2 Seizure Data Set

The seizure data set was analysed in order to discern whether pre-ictal and ictal states of the test person could be differentiated through iEEG signals.

13.2.1 All Electrodes

As seen in Figures 12.15 and 12.16 the results from the seizure data set showed a significant difference between the results from multiple dependency measures and signal representations confirmed by the p-values from Welch's t-test in Table 12.1. In particular $\Omega(\rho_c; s_p)$ and all OCs based upon MI showed significant results along with O-information of all representations. The results therefore suggest, that it is possible to differentiate pre-ictal and ictal states through iEEG signals and these dependency measures on all electrodes. Notice furthermore that all OCs based upon MI and $\mathcal{O}(s_p)$ from visual inspection seem to display different variances for pre-ictal and ictal results, indicating that a potential differentiation can be made through likelihoods and posteriors if these variances were estimated.

The data set contains signals from 76 electrodes which results in $\sum_{i=1}^{76} i = 2926$ bivariate dependencies. If some of the OCs can differentiate between pre-ictal and ictal states through EEG signals with 76 electrodes, it indicates that there is a large enough difference between the EEG signals to causes a noteworthy difference in these bivariate dependencies.

Notice however that the OCs in Figure 12.15 are very small, and that it is assumed that the large number of electrodes is the cause of particularly small OCs.

Since only the OCs based upon CCC and MI produce significant results, this suggests that these dependency functions capture differences between dependencies in the pre-ictal and ictal data sets better than PCC and that these are more suited for analysis of this type of signals. It furthermore suggests that the differences between pre-ictal and ictal data sets are more easily described through phase information or non-linear dependencies.

The O-information is positive for both pre-ictal and ictal states, indicating redundancy dominated data sets. The O-information is furthermore for s_r and s_a significantly higher for pre-ictal states indicating stronger redundant dependencies.

13.2.2 Clustered Data Set

The clusterings seen in Figures 12.17-12.19 are similar but not equal. The cluster c_{red} in the bottom of the electrode configurations in particular is almost repeated. The overlap between clusters suggests noteworthy communities.

When examining results from the clusterings c_{PCC} and c_{MI}, multiple OCs and O-information were significant as seen in Figures 12.20-12.23 and confirmed by Tables 12.2 and 12.3. In particular the red clusters $c_{PCC,red}$ and $c_{MI,red}$ seem to contain information relevant for discriminating pre-ictal and ictal states. This is clear since both OCs and both O-information yield significant results. This observation is consistent with the observations made in Kramer et al. [2008].

The p-values of Welch's t-test of the results from applying all dependency measures and signals representations on the clustering c_{TE} are shown in Table 12.4. The OCs $\Omega(\rho_p; s_r)$ and $\Omega(\rho_p; s_p)$ produce significant results in most clusters, that is in clusters $c_{TE,red}$, $c_{TE,blue}$ and $c_{TE,green}$, suggesting that these best capture the differences between pre-ictal and ictal states based upon the clustering c_{TE}. It is furthermore clear that all dependency measures except $\Omega(\rho_p; s_a)$ and $\Omega(\rho_c; s_p)$ applied on cluster $c_{TE,red}$ produce significant results suggesting that $c_{TE,red}$ is suitable for differentiating between pre-ictal and ictal states. The p-values for $c_{TE,red}$ are in addition particularly small for $\Omega(I; s_r)$, $\Omega(I; s_a)$, $\Omega(I; s_p)$, $\mathcal{O}(s_r)$, $\mathcal{O}(s_a)$ and $\mathcal{O}(s_p)$.

In addition to the significant differences in mean values, some sets of results from visual inspection show apparent differences in variances. This can in particular be seen from $\mathcal{O}(s_r)$ on $c_{PCC,blue}$ in Figure 12.21, $\Omega(\rho_p; s_a)$ on $c_{TE,purple}$ in Figure 12.25 and $\Omega(I; s_p)$ on $c_{TE,red}$ in Figure 12.30 and suggests that a potential differentiation between pre-ictal and ictal can be made through likelihoods and posteriors if the variances of the results were estimated.

Notice that all results based upon clustering can again be prone to survivorship bias.

The pre-ictal and ictal states could be discriminated without clustering but the clusterings $c_{PCC,red}$, $c_{MI,red}$ and $c_{TE,red}$ facilitated a better discrimination between the two states.

13.2.3 Relation to SNR

Based on the results from the seizure data set, it is clear that some of the combinations of dependency measures and signal representations in Section 10.2 might be useful for discrimination between pre-ictal and ictal states a of subject based upon iEEG signals. This stimulates the hope that some of the dependency measures and signal representations in conjunction with clustering of the electrodes can lead to a discrimination between EEG signals collected during exposure to different stimuli, for example various signal to noise ratios.

14 | Conclusion

Motivated by research suggesting that utilising analysis of dependencies in electroencephalography (EEG) signals can be used for a deeper understanding of the activity of the brain in different circumstances and with the goal of finding correlates in EEG signals of listening effort a number of mathematical tools for describing multivariate dependencies were established. This included a generalisation of the omega complexity and information theoretic quantities. These were used in conjunction with graph theory in a data driven approach seeking to find subsets of sensors in EEG signals which more clearly show differences in activity of the brain.

The introduced tools were applied to simulated Rössler systems and multivariate autoregressive processes, serving as an indication that the introduced methods could be applicable for detecting dependencies in EEG data, since they produced different results reflecting changes in coupling configurations and degrees.

Results from analysis of an EEG data set obtained from subjects performing tasks at different levels of SNR showed no significant differences in the means of results from comparable experiments.

Results from analysis of an intrusive EEG (iEEG) data set from pre-ictal and ictal states of an epileptic person showed significant differences between multiple results and with more pronounced significance after clustering. One cluster in particular provided the basis for multiple significant results suggesting that certain subsets of iEEG electrodes are better for registering differences between signals from pre-ictal and ictal states.

The results obtained from simulated signals in conjunction with the results obtained from real signals showed that the mathematical tools to some degree capture different dependencies in different signals and that these might be useful for quantifying dependencies in EEG signals.

This project has as such through mathematical analysis, examples and results from simulated and real signals conducted an examination of how the omega complexity, information theory and graph theory may be used to quantify dependencies in EEG signals for analysis of the activity of the human brain, while also showing deficiencies of the tools applied and motivating further work into both theoretical and practical aspects of the process of analysing EEG signals.

14.1 Future work

The following points could be considered for future work:

- **Neurophysiological Perspective**
  The analysis of EEG signals could be aided with neurophysiological knowledge. This could give directions for the methodology with knowledge of how the nature of EGG signals might change under different conditions. Furthermore the results regarding subsets of EEG electrodes could have been related to sections of the brain derived from neurophysiological theory, which might be relevant for an analysis seeking to quantify dependencies.

- **Additional Data**
  Analysing additional data sets could aid the conclusions of the project. More specifically it could help detailing if different changes in conditions result in different changes of dependencies.

- **Statistical Tests**
  The significance of the results relies solely on Welch's t-test. Other statistical tests could be made in order to examine whether the results contain any additional meaningful insights regarding activity in the brain.

Bibliography

Ala, T. S., Graversen, C., Wendt, D., Alickovic, E., Winther, W. M., and Lunner, T. (2022). An exploratory study of eeg alpha oscillation and pupil dilation in hearing-aid users during effortful listening to continuous speech. *Ear and Hearing*.

Almerida, H., Guedes, D., Jr., W. M., and Zaki, M. J. (2019). Is there a best quality metric for graph clusters? `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.228.3004&rep=rep1&type=pdf`. Date accessed: 31/05-2023.

Ameera, A., A.Saidatul, and Ibrahim, Z. (2018). Analysis of eeg spectrum bands using power spectral density for pleasure and displeasure state.

Atyabi, A., Shic, F., and Naples, A. (2016). Mixture of autoregressive modeling orders and its implication on single trial eeg classification. *Expert Systems with Appplications*.

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence.

Baboukani, P. S., Azemi, G., Boashash, B., and Omidvarnia, P. C. A. (2018). A novel multivariate phase synchrony measure: Application to multichannel newborn eeg analysis. *Digital Signal Processing*.

Baboukani, P. S., Graversen, C., Alickovic, E., and Østergaard, J. (2020). Estimating conditional transfer entropy in time series using mutual information and nonlinear prediction. *MDPI*.

Baboukani, P. S., Graversen, C., Alickovic, E., and Østergaard, J. (2022). Speech to noise ratio improvement induces nonlinear parietal phase synchrony in hearing aid users. *Frontiers*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) , 1995, Vol. 57, No. 1 (1995), pp. 289-300*.

Biasiucci, A., Franceschiello, B., and Murray, M. M. (2019). Electroencephalography. *Current Biologi*.

Bondy, J. A. and Murty, U. S. R. (1982). *Graph Theory with Applications*. Elsevier.

Bondy, J. A. and Murty, U. S. R. (2008). *Graph Theory*. Springer.

Britton, J. W., Frey, L. C., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., Pestana-Knight, E. M., and and, E. K. S. L. (2016). *Electroencephalography: An introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*.

Cover, T. M. and Thomas, J. A. (2016). *Elemensents of Information Theory*. John Wiley and Sons.

Diestel, R. (2017). *Graph Theory*. Springer, 5 edition.

Faes, L. (2019). Ts - matlab tool for the computation of information dynamics.

Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika , Aug., 1983, Vol. 70, No. 2 (Aug., 1983), pp. 327-332*.

Gao, S., Steeg, G. V., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. *arXiv:1411.2003*.

Gaál, Z. A., Boha, R., Stam, C. J., and Molnár, M. (2010). Age-dependent features of eeg-reactivity – spectral, complexity and network characteristics. *Neuroscience Letters*.

Hjortkjær, J., Märcher-Rørsted, J., Fuglsang, S. A., and Dau, T. (2020). Cortical oscillations and entrainment in speech processing during working memory load. *European Journal of Neuroscience*.

Irisawa, S., Isotani, T., Yagyu, T., Morita, S., Nishida, K., Yamada, K., Yoshimura, M., Okugawa, G., Nobuhara, K., and Kinoshita, T. (2006). Increased omega complexity and decreased microstate duration in nonmedicated schizophrenic patients. *Neuropsychobiology 2006;54:134–139*.

Jakobsen, S. K. (2013). Mutual information matrices are not always positive semi-definite.

Jia, H., Li, Y., and Yu, D. (2018). Normalized spatial complexity analysis of neural signals. *Scientific Reports*.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., III, D. J. E., Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*.

Kramer, M. A., Kolaczyk, E. D., and Kirsch, H. E. (2008). Emergent network topology at seizure onset in humans. *Epilepsy Research*.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E 69.6*.

Kschischang, F. R. (2015). The hilbert transform.

Liu, Y.-W. (2012). Hilbert transform and applications. *Fourier Transform Applications*, pages 291–300.

Marple, S. L. (1999). Computing the discrete-time "analytic" signal via fft. *IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 47, NO. 9*.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., and Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? a british society of audiology cognition in hearing special interest group ' white paper '. *Informa Healthcare*.

Mensen, A., Marshall, W., and Tononi, G. (2017). Eeg differentiation analysis and stimulus set meaningfulness. *Frontier in Psychology*. `https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01748/full`.

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., and Lyxell, B. (2017). Objective assessmensent of listening effort: Coregistration of pupillometry and eeg. *Trends in Hearing*.

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in meg/eeg: a review and recommensendations. *frontiers in Human Neuroscience*.

NCSS (2023). Ncss data analysis & graphs.

Newman, M. and Girvan, M. (August 2003). Finding and evaluating community structure in networks. `https://arxiv.org/abs/cond-mat/0308217`.

Nielsen, R. M. and Ng, E. H. N. (2022). Oticon more™ new evidence —- reducing sustained listening effort.

Oh, W. G. S. and Viswanath, P. (2016). Demystifying fixed k-nearest neighbor information estimators.

Oticon (2023). It's time for the new perspective in brainhearing™.

Peele, J. E. (2017). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*.

Pounder, K. J. and Sauer, D. T. D. (2009). Synchronization and coherence of dynamical systems: Networks of coupled rössler attractors.

Rosas, F., Mediano, P. A., Gastpar, M., and Jensen, H. J. (2019). Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E 100*.

Sakai, T. (2016). Two sample t-tests for ir evaluation: Student or welch? *SIGIR '16, July 17 - 21, 2016*.

Santurette, S., Ng, E. H. N., Jensen, J. J., and Loong, B. M. K. (2020). Oticon more™ clinical evidence.

Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett. 85, 461*.

Serre, D. (2010). *Matrices: Theory and Applications, Second Edition*. Springer.

Starn, C., van derLeij, E., Meulstee, J., and Vliegen, J. H. (2000). Changes in functional coupling between neural networks in the brain during maturation revealed by omega complexity. *Clinical Electroencephalography*.

Subramaniyam, N. P. and Hyttinen, J. (2014). Characterization of dynamical systems under noise using recurrence networks: Application to simulated and eeg data. *Physics Letter A V.378, I. 46*.

Toth, M., Faludi, B., Wackermann, J., Czopf, J., and Kondakor, I. (2009). Characteristic changes in brain electrical activity due to chronic hypoxia in patients with obstructive sleep apnea syndrome (osas): A combined eeg study using loreta and omega complexity. *Brain Topography 22, p. 185–190 (2009)*.

Tsipouras, M. G. (2019). Spectral information of eeg signals with respect to epilepsy classification.

University of California San Fransisco (2023). Hearing loss. Accessed: 30-05-2023.

Wackermann, J. (1996). Beyond mapping: estimating complexity of multichannel eeg recordings. *Acta Neurobiologiae Experimensentalis*.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika, Vol. 34, No. 1/2 (Jan., 1947), pp. 28-35 (8 pages)*.

WHO (2007). Nearly 1 in 6 of world's population suffer from neurological disorders. `https://news.un.org/en/story/2007/02/210312-nearly-1-6-worlds-population-suffer-neurological-disorders-un-report`. Accessed: 22-02-2022.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv*.

World Health Organization (2023). Deafness and hearing loss. Accessed: 30-05-2023.

Xefteris, V.-R., Tsanousa, A., Georgakopoulou, N., Sotiris Diplaris, S. V., and Kompatsiaris, I. (2022). Graph theoretical analysis of eeg functional onnectivity patterns and fusion with physiological signals for emotion recognition. *Sensors*.

Zhang, Y., Liu, B., Ji, X., and Huang, D. (2018a). Classification of eeg signals based on autoregressive model and wavelet packet decomposition. *Neural Processing Letters*.

Zhang, Y., Zhang, S., and Ji, X. (2018b). Eeg-based classification of emotions using empirical mode decomposition and autoregressive model. *Multimedia Tools and Applications*.

Zhao, S., Ng, S.-C., Khoo, S., and Chi, A. (2022). Temporal and spatial dynamics of eeg features in female college students with subclinical depression. *International Journal of Environmensental Research and Public Health*.

Appendices

A | Phase Shifted Sine Waves

The sets $S_1, \ldots, S_4$ each consisting of three phase shifted sine waves and which are used in Section 7.1.2 are defined as follows:

$$S_1 = \begin{Bmatrix} s_{1,1}(t) & = \sin(t), \\ s_{1,2}(t) & = \sin(t + 2\pi/3), \\ s_{1,3}(t) & = \sin(t + 4\pi/3) \end{Bmatrix} \qquad S_2 = \begin{Bmatrix} s_{2,1}(t) & = \sin(t), \\ s_{2,2}(t) & = \sin(t + 2\pi/3), \\ s_{2,3}(t) & = \sin(t + 4\pi/3) \end{Bmatrix}$$

$$S_3 = \begin{Bmatrix} s_{3,1}(t) & = \sin(t), \\ s_{3,2}(t) & = \sin(t + 2\pi/3), \\ s_{3,3}(t) & = \sin(t + 4\pi/3) \end{Bmatrix} \qquad S_4 = \begin{Bmatrix} s_{4,1}(t) & = \sin(t), \\ s_{4,2}(t) & = \sin(t + 2\pi/3), \\ s_{4,3}(t) & = \sin(t + 4\pi/3) \end{Bmatrix}$$

B | Rössler Transfer Entropy Results

In this appendix, all transfer entropies of Rössler systems are shown.

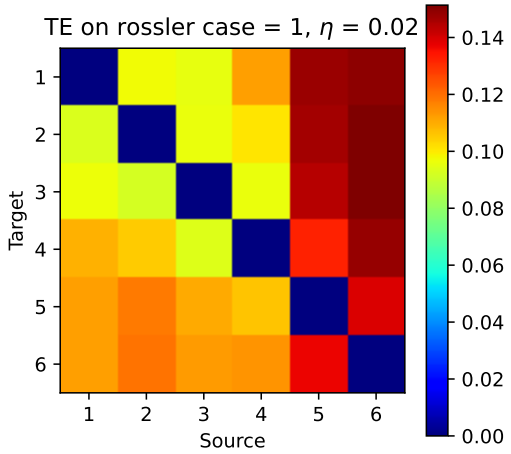


Figure B.1: TE on Rössler Case 1, $\eta = 0.02$

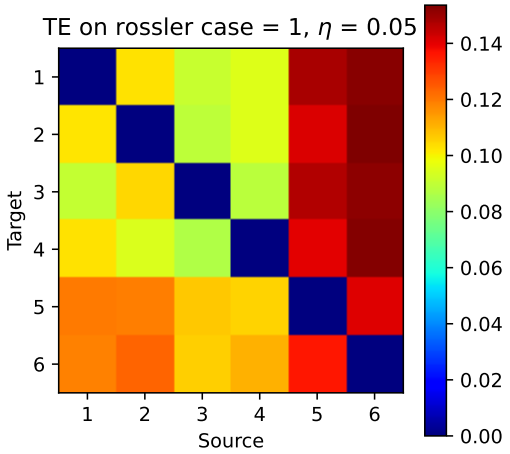


Figure B.2: TE on Rössler Case 1, $\eta = 0.05$

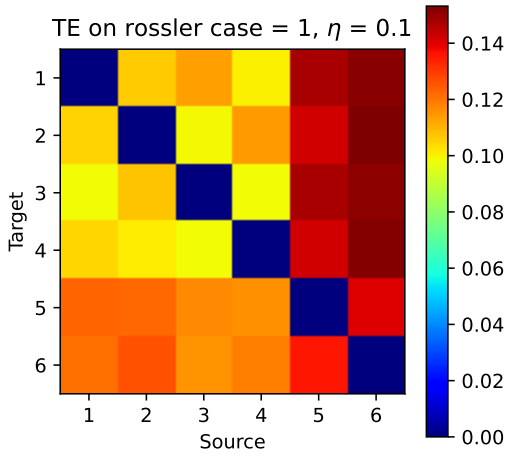


Figure B.3: TE on Rössler Case 1, $\eta = 0.1$

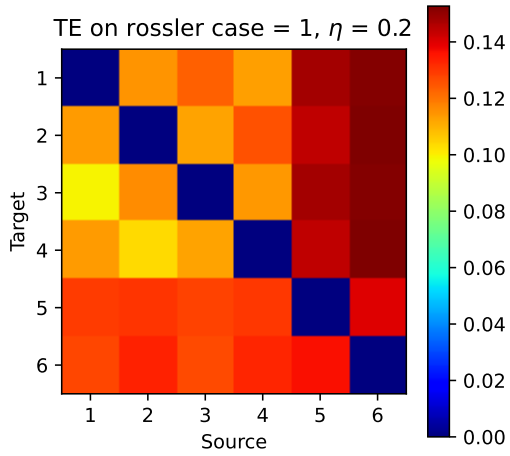


Figure B.4: TE on Rössler Case 1, $\eta = 0.2$

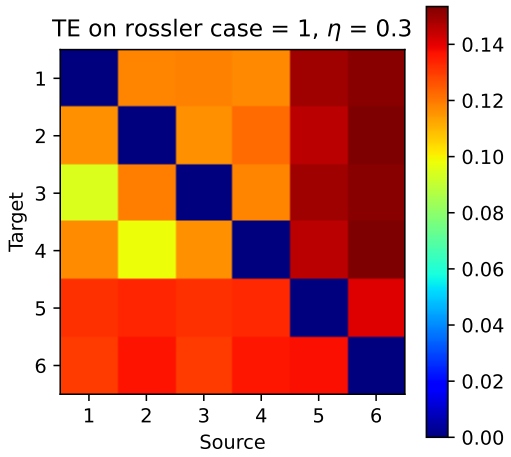


Figure B.5: TE on Rössler Case 1, $\eta = 0.3$

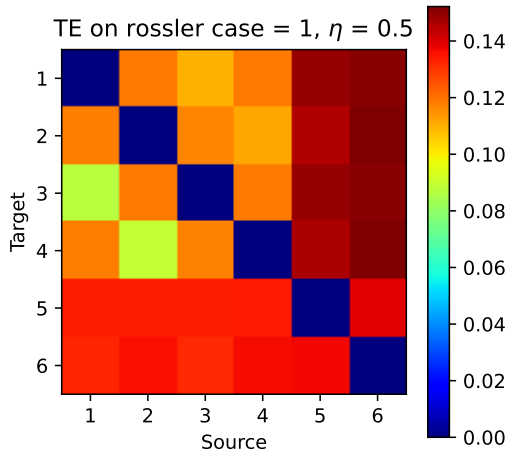


Figure B.6: TE on Rössler Case 1, $\eta = 0.5$

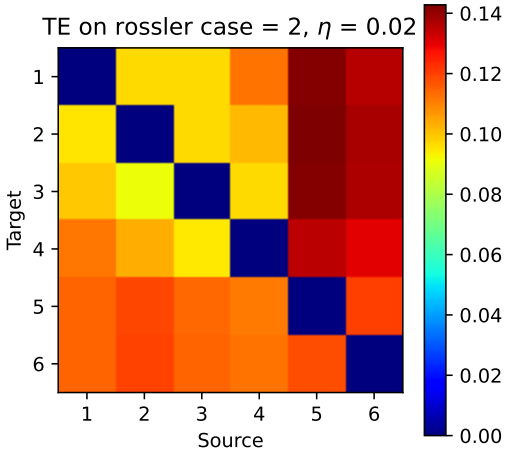


Figure B.7: TE on Rössler Case 2, $\eta = 0.02$

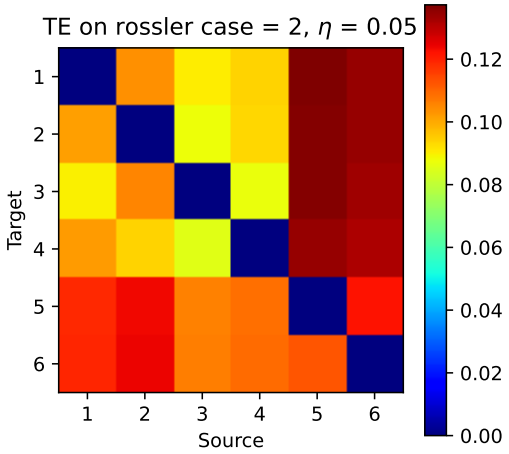


Figure B.8: TE on Rössler Case 2, $\eta = 0.05$

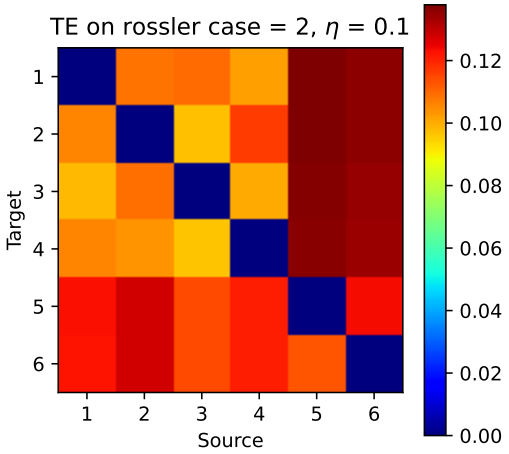


Figure B.9: TE on Rössler Case 2, $\eta = 0.1$

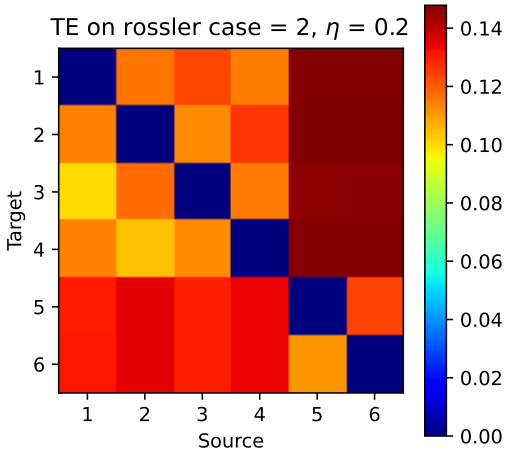


Figure B.10: TE on Rössler Case 2, $\eta = 0.2$

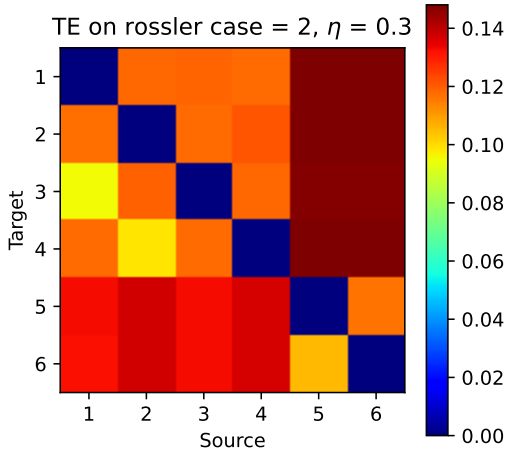


Figure B.11: TE on Rössler Case 2, $\eta = 0.3$

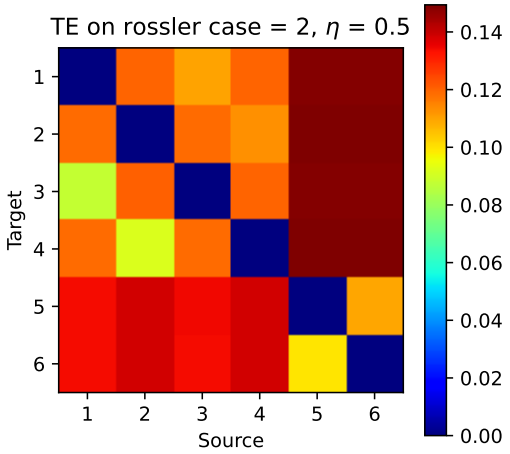


Figure B.12: TE on Rössler Case 1, $\eta = 0.5$

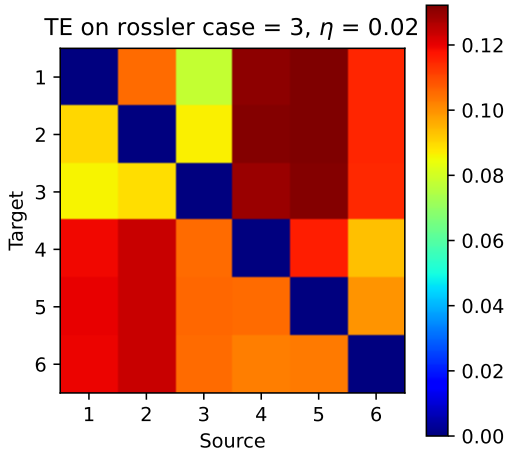


Figure B.13: TE on Rössler Case 3, $\eta = 0.02$

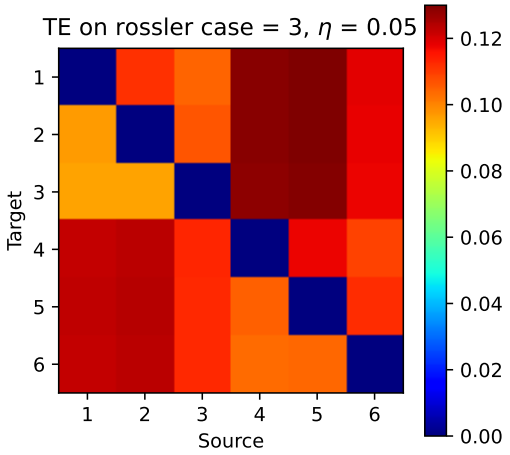


Figure B.14: TE on Rössler Case 3, $\eta = 0.05$

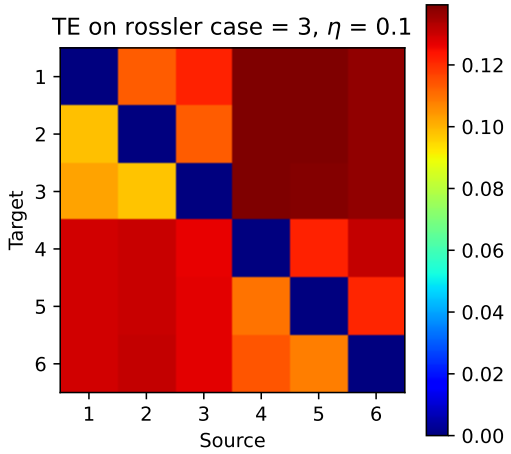


Figure B.15: TE on Rössler Case 3, $\eta = 0.1$

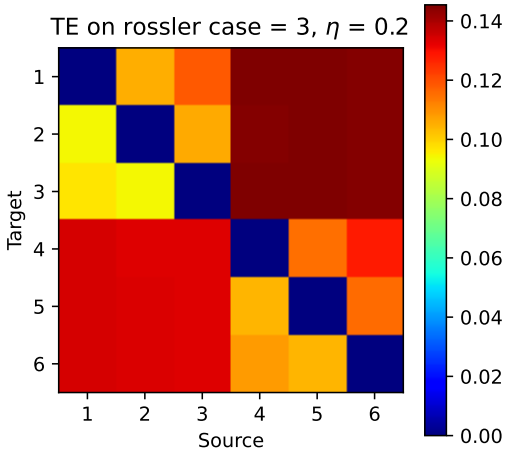


Figure B.16: TE on Rössler Case 3, $\eta = 0.2$

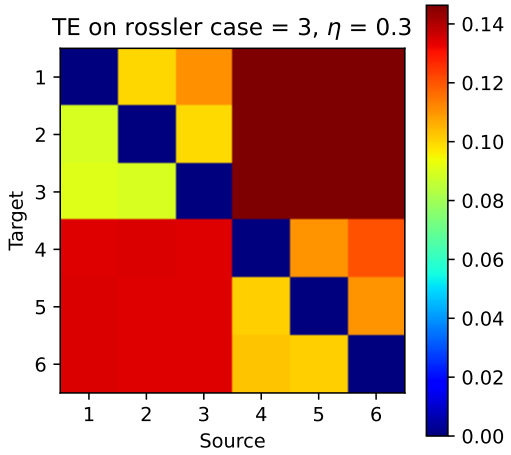


Figure B.17: TE on Rössler Case 3, $\eta = 0.3$

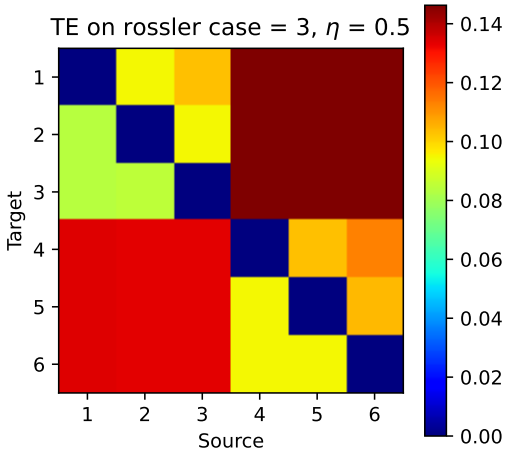


Figure B.18: TE on Rössler Case 1, $\eta = 0.5$

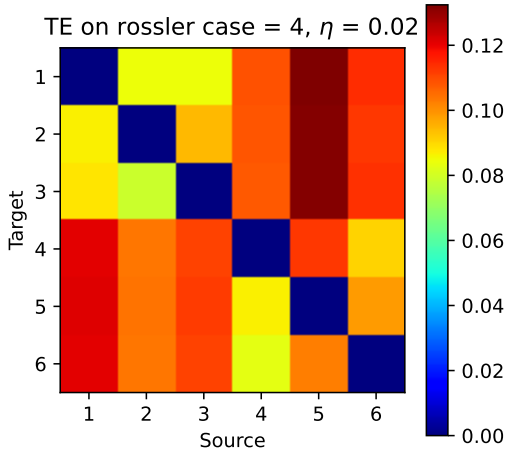


Figure B.19: TE on Rössler Case 4, $\eta = 0.02$

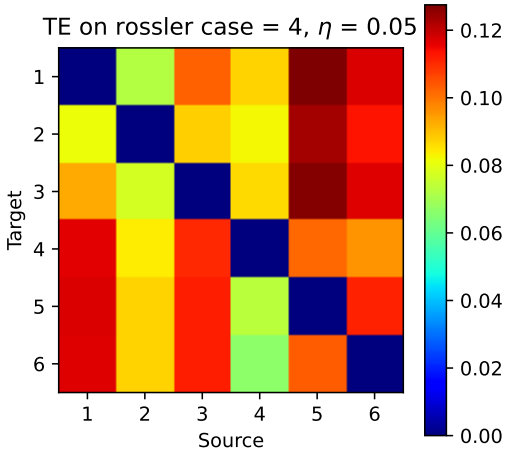


Figure B.20: TE on Rössler Case 4, $\eta = 0.05$

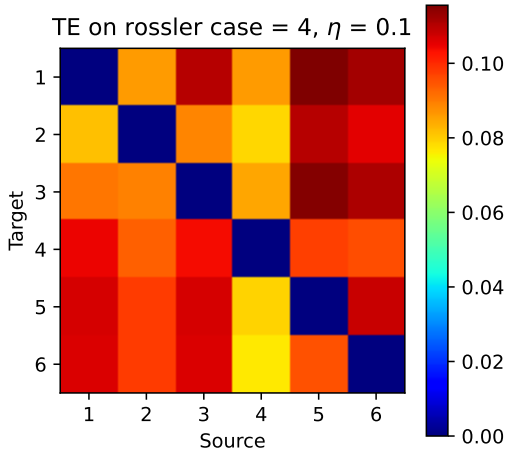


Figure B.21: TE on Rössler Case 4, $\eta = 0.1$

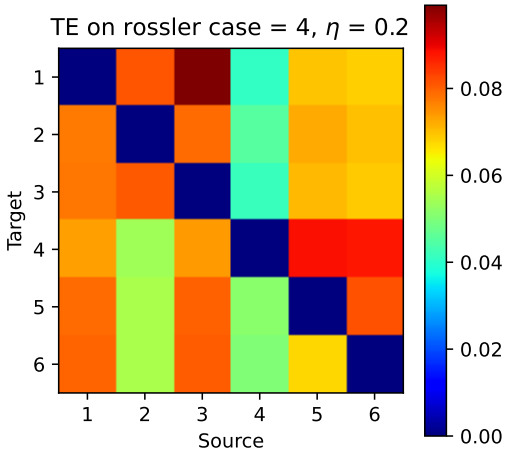


Figure B.22: TE on Rössler Case 4, $\eta = 0.2$

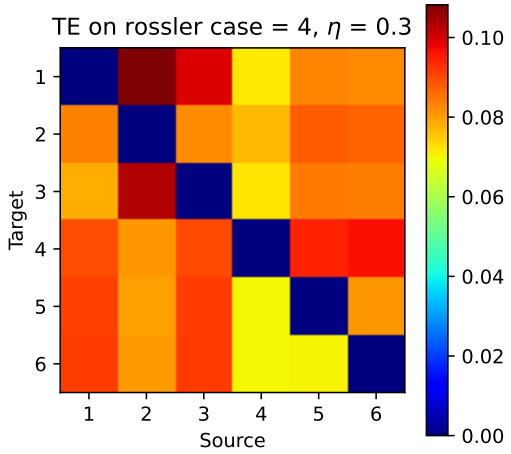


Figure B.23: TE on Rössler Case 4, $\eta = 0.3$

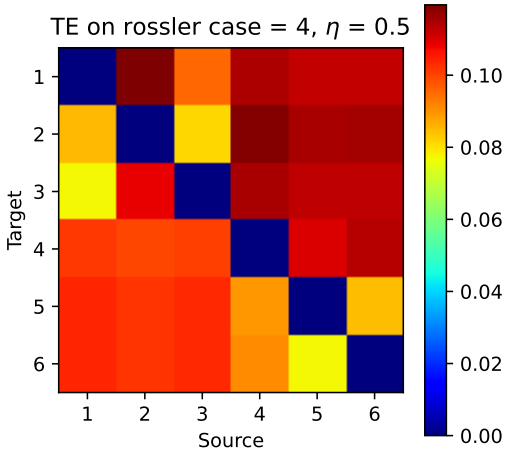


Figure B.24: TE on Rössler Case 4, $\eta = 0.5$

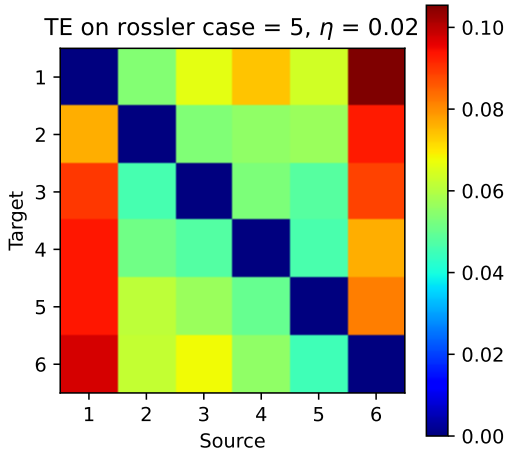


Figure B.25: TE on Rössler Case 5, $\eta = 0.02$

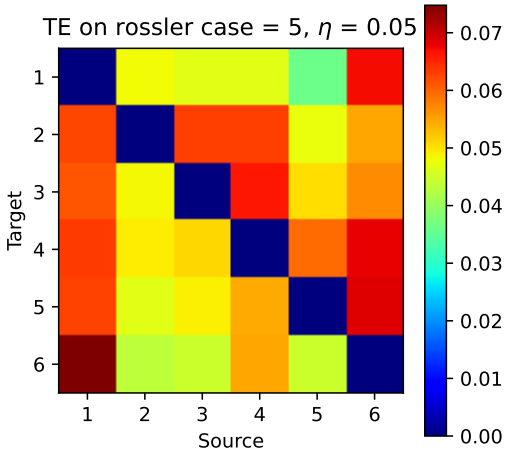


Figure B.26: TE on Rössler Case 5, $\eta = 0.05$

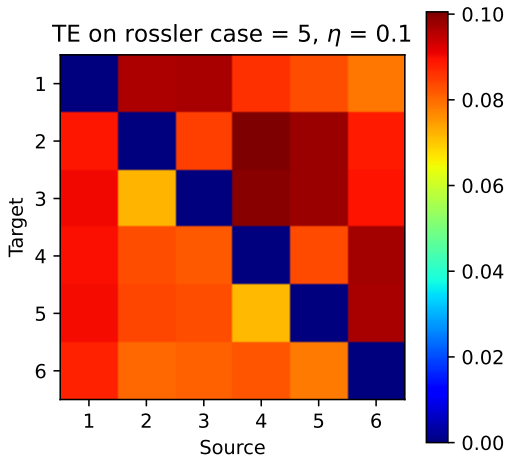


Figure B.27: TE on Rössler Case 5, $\eta = 0.1$

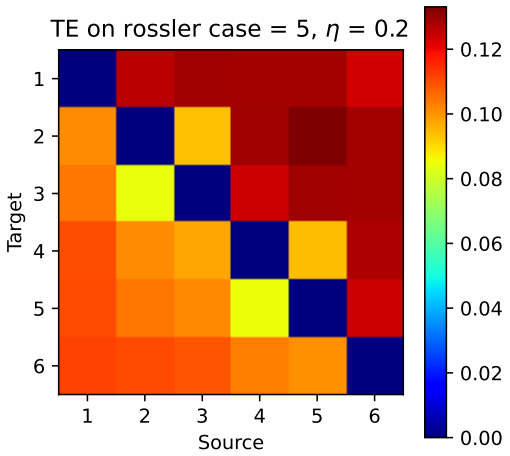


Figure B.28: TE on Rössler Case 5, $\eta = 0.2$

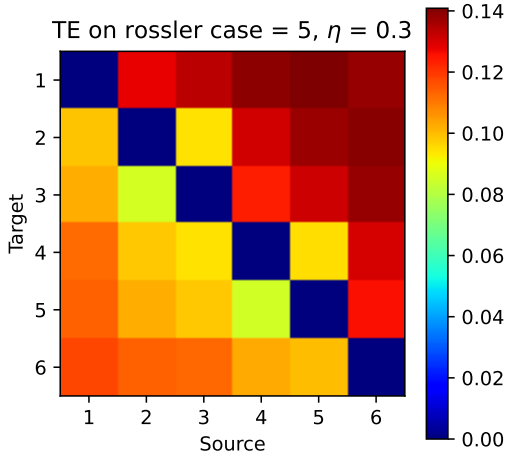


Figure B.29: TE on Rössler Case 5, $\eta = 0.3$

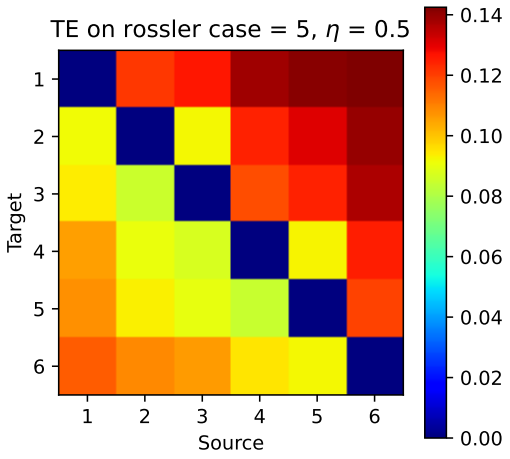


Figure B.30: TE on Rössler Case 5, $\eta = 0.5$

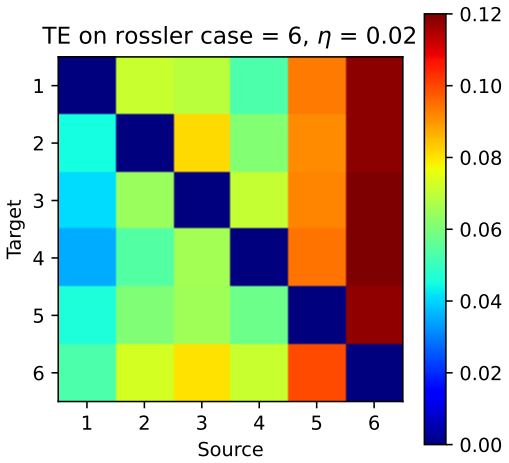


Figure B.31: TE on Rössler Case 6, $\eta = 0.02$

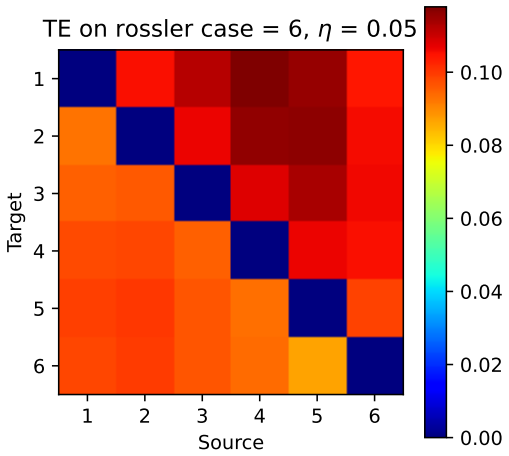


Figure B.32: TE on Rössler Case 6, $\eta = 0.05$

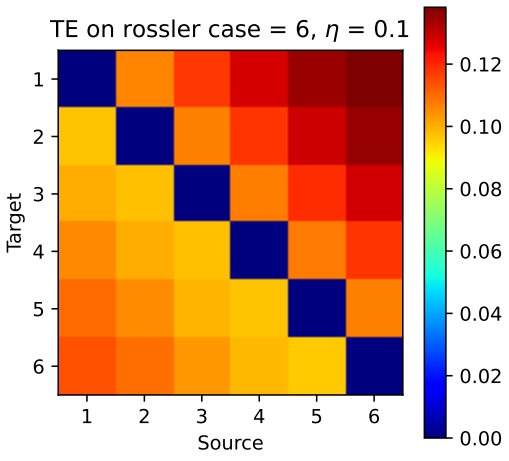


Figure B.33: TE on Rössler Case 6, $\eta = 0.1$

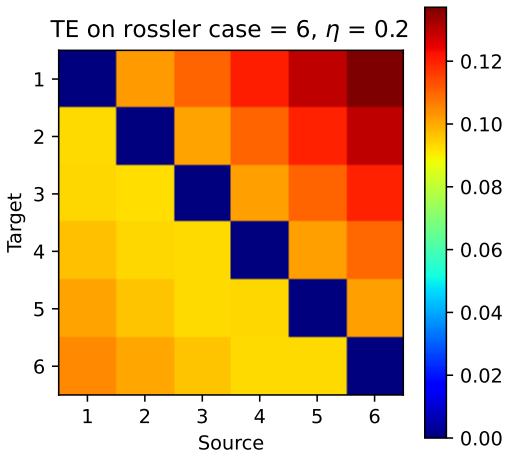


Figure B.34: TE on Rössler Case 6, $\eta = 0.2$

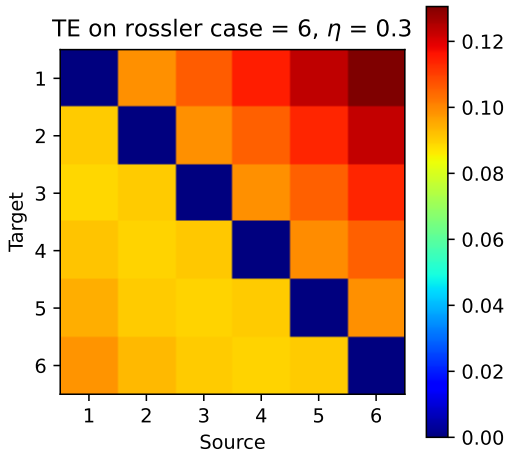


Figure B.35: TE on Rössler Case 6, $\eta = 0.3$

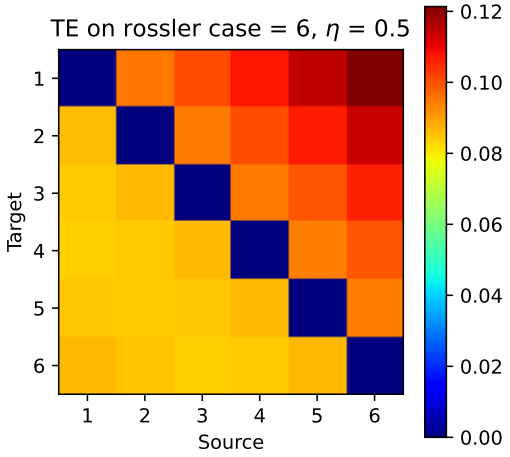


Figure B.36: TE on Rössler Case 6, $\eta = 0.5$

C | Clustering MVAR TE Results

This appendix contains graphs for all αs in MVAR with edges colored by transfer entropy and nodes colored by cluster found by the Louvain algorithm.

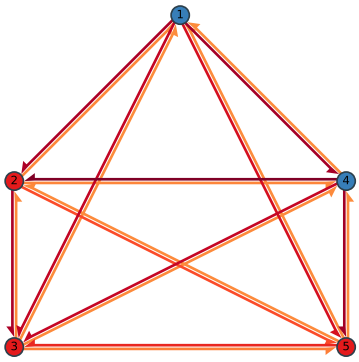


Figure C.1: Clusterings with on MVAR with $\alpha = 0$.

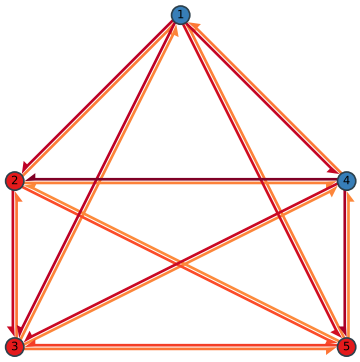


Figure C.2: Clusterings with on MVAR with $\alpha = 0.1225$.

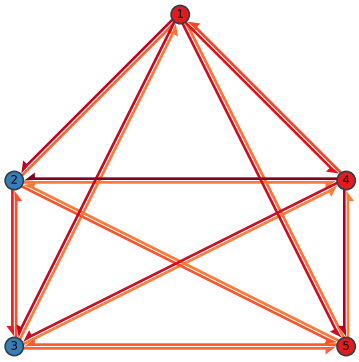


Figure C.3: Clusterings with on MVAR with $\alpha = 0.245$.

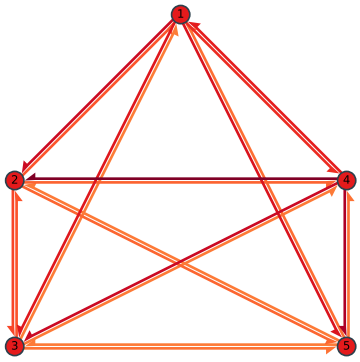


Figure C.4: Clusterings with on MVAR with $\alpha = 0.3657$.

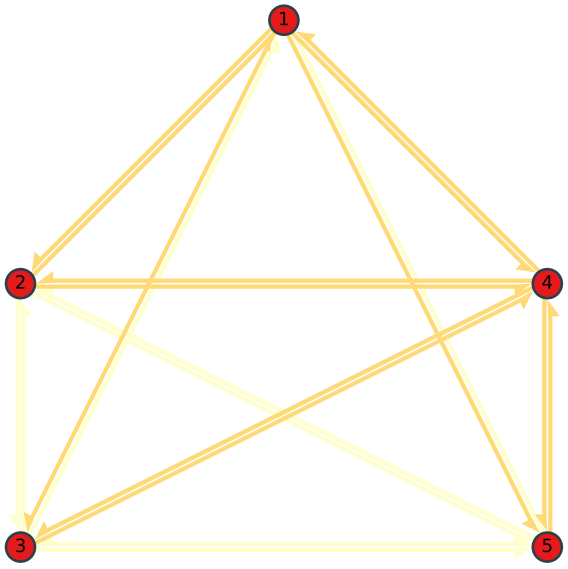


Figure C.5: Clusterings with on MVAR with $\alpha = 0.49$.

D | EEG Results

This appendix contains all results of omega complexity and O-information for the 18 subjects analysed in Chapter 12.

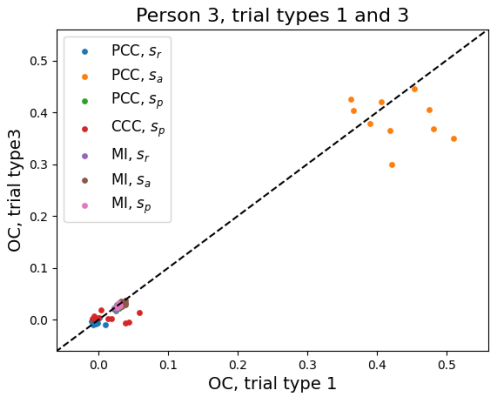


Figure D.1: OCs of all repeats of trials 1 and 3 for person 3.

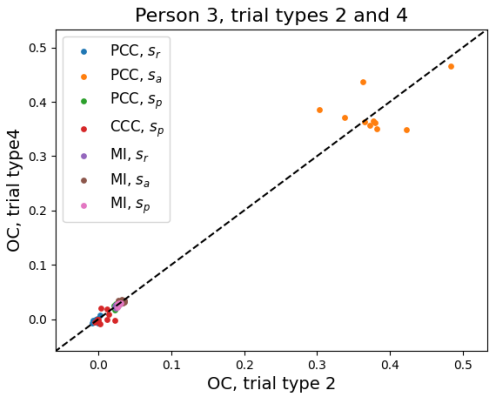


Figure D.2: OCs of all repeats of trial type 2 and 4 for person 3.

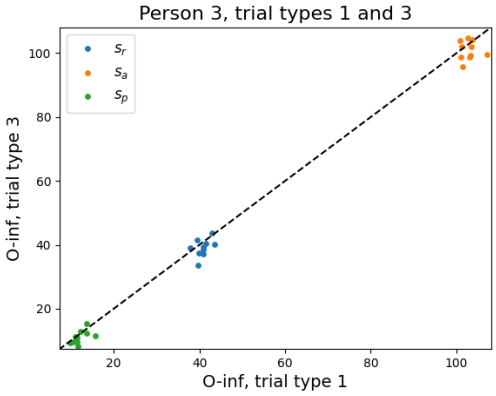


Figure D.3: O-information of all repeats of trial type 1 and 3 for person 3.

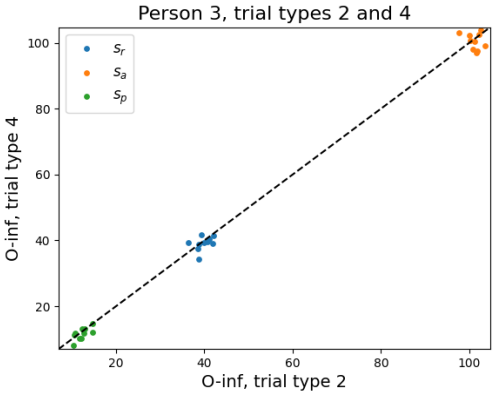


Figure D.4: O-information of all repeats of trial type 2 and 4 for person 3.

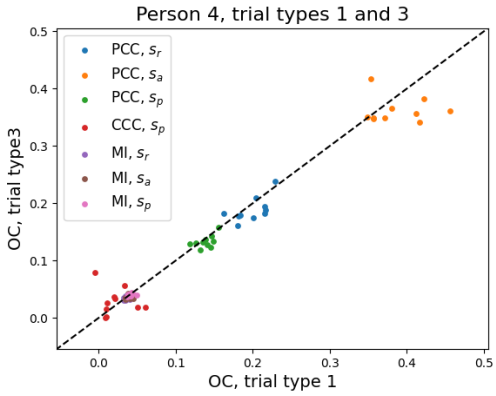


Figure D.5: OCs of all repeats of trials 1 and 3 for person 4.

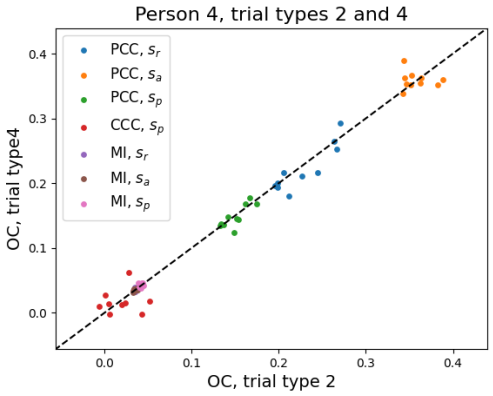


Figure D.6: OCs of all repeats of trial type 2 and 4 for person 4.

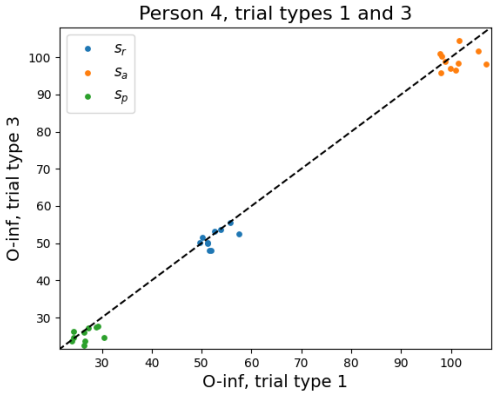


Figure D.7: O-information of all repeats of trial type 1 and 3 for person 4.

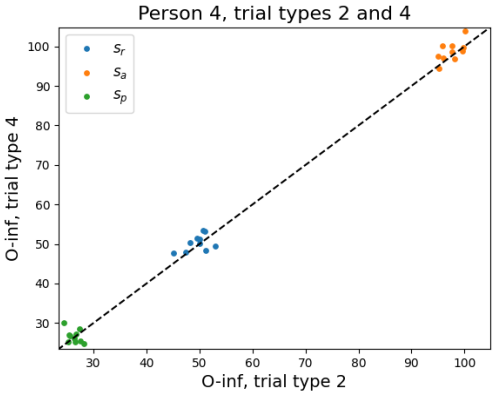


Figure D.8: O-information of all repeats of trial type 2 and 4 for person 4.

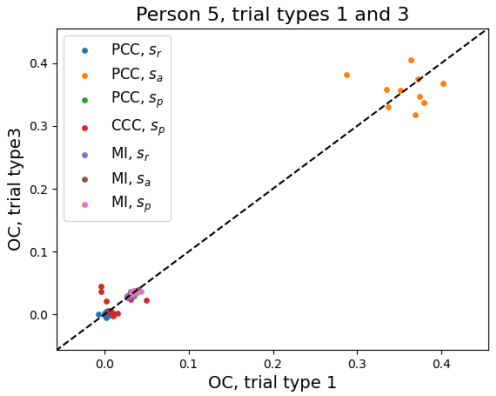


Figure D.9: OCs of all repeats of trials 1 and 3 for person 5.

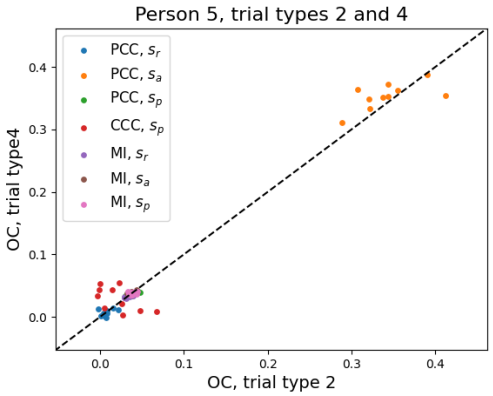


Figure D.10: OCs of all repeats of trial type 2 and 4 for person 5.

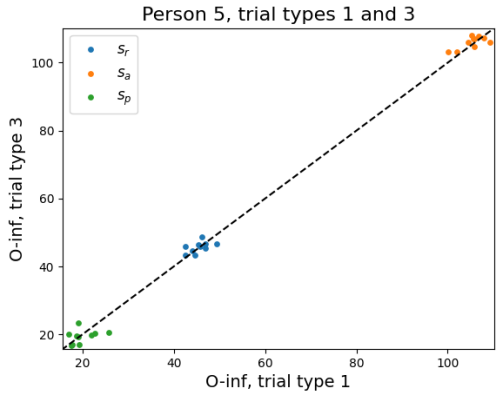


Figure D.11: O-information of all repeats of trial type 1 and 3 for person 5.

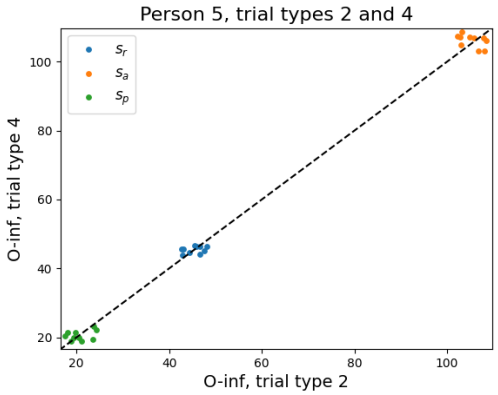


Figure D.12: O-information of all repeats of trial type 2 and 4 for person 5.

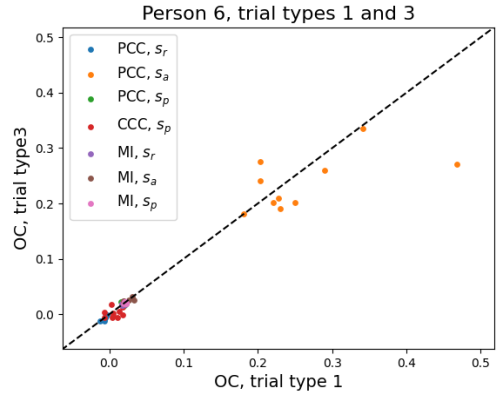


Figure D.13: OCs of all repeats of trials 1 and 3 for person 6.

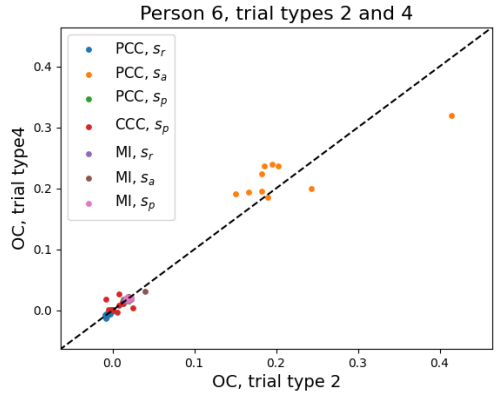


Figure D.14: OCs of all repeats of trial type 2 and 4 for person 6.

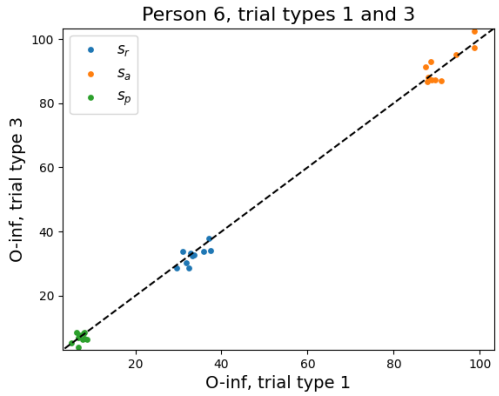


Figure D.15: O-information of all repeats of trial type 1 and 3 for person 6.

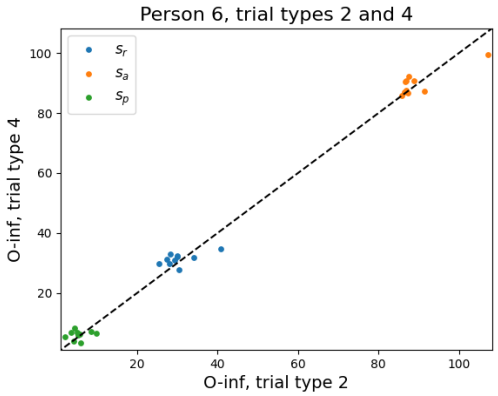


Figure D.16: O-information of all repeats of trial type 2 and 4 for person 6.

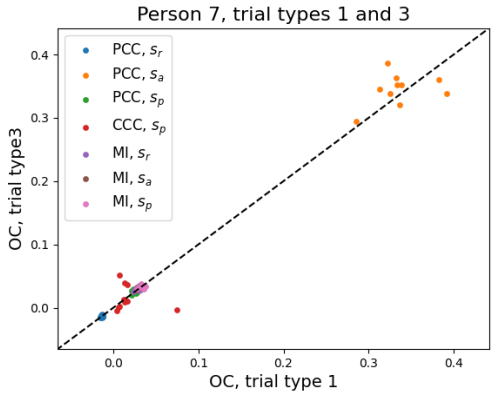


Figure D.17: OCs of all repeats of trials 1 and 3 for person 7.

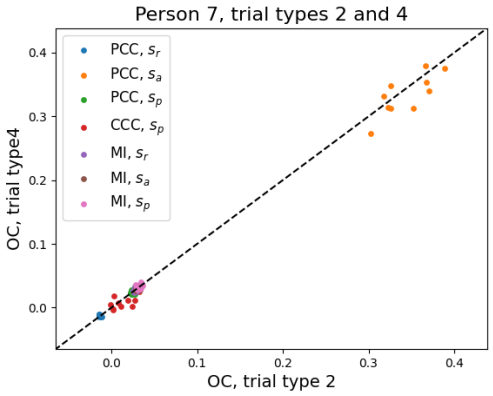


Figure D.18: OCs of all repeats of trial type 2 and 4 for person 7.

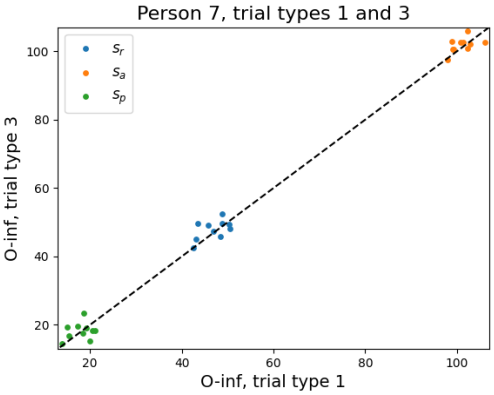


Figure D.19: O-information of all repeats of trial type 1 and 3 for person 7.

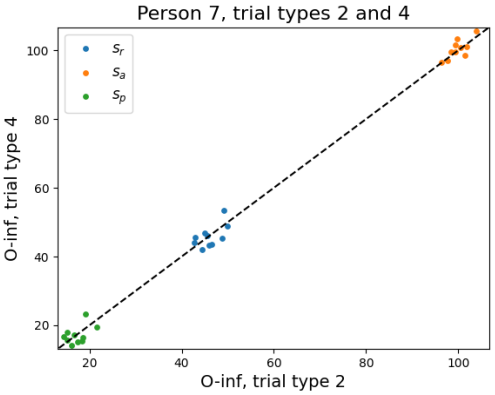


Figure D.20: O-information of all repeats of trial type 2 and 4 for person 7.

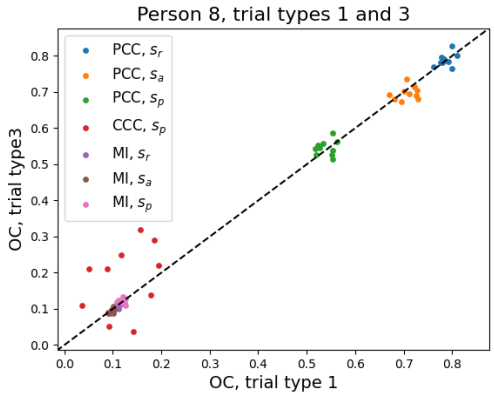


Figure D.21: OCs of all repeats of trials 1 and 3 for person 8.

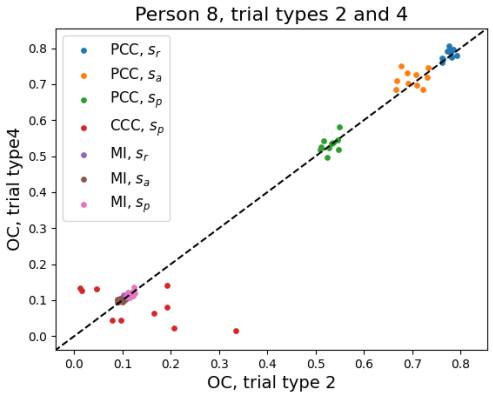


Figure D.22: OCs of all repeats of trial type 2 and 4 for person 8.

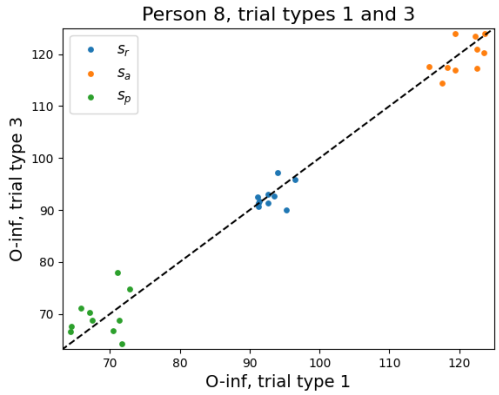


Figure D.23: O-information of all repeats of trial type 1 and 3 for person 8.

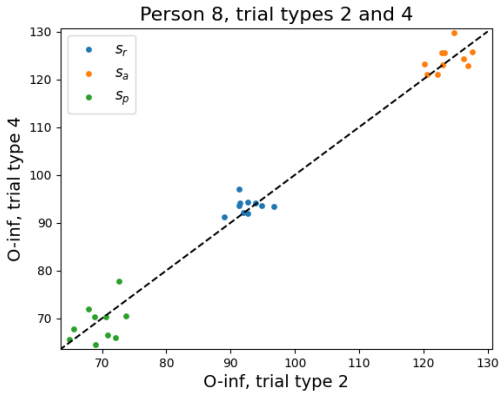


Figure D.24: O-information of all repeats of trial type 2 and 4 for person 8.

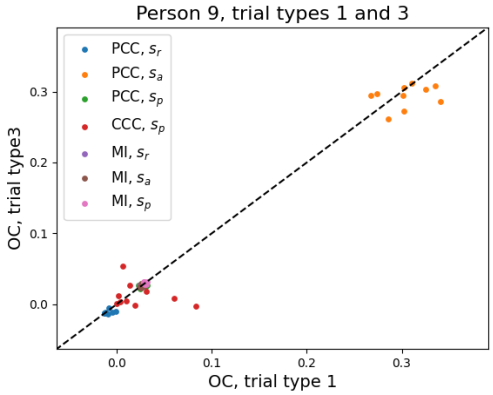


Figure D.25: OCs of all repeats of trials 1 and 3 for person 9.

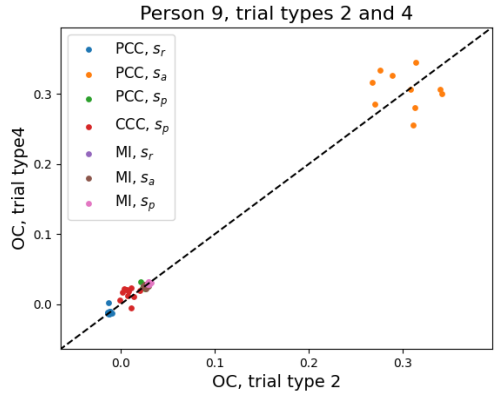


Figure D.26: OCs of all repeats of trial type 2 and 4 for person 9.

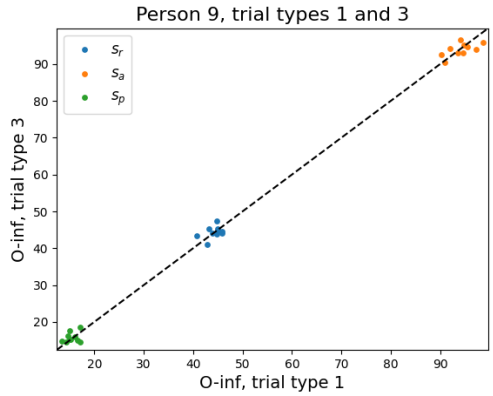


Figure D.27: O-information of all repeats of trial type 1 and 3 for person 9.

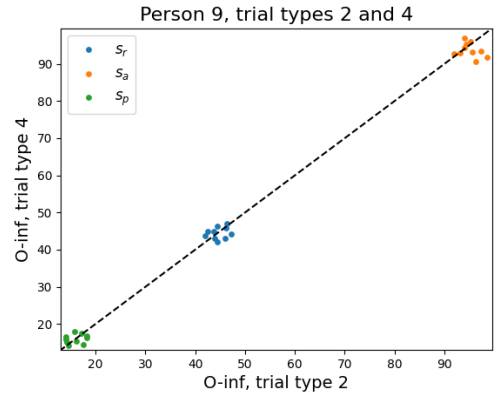


Figure D.28: O-information of all repeats of trial type 2 and 4 for person 9.

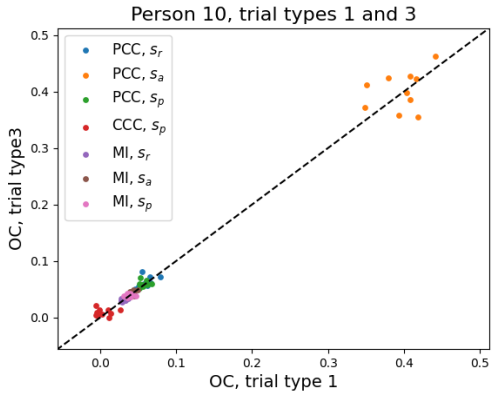


Figure D.29: OCs of all repeats of trials 1 and 3 for person 10.

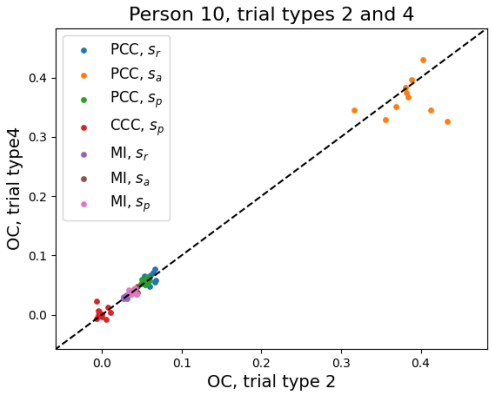


Figure D.30: OCs of all repeats of trial type 2 and 4 for person 10.

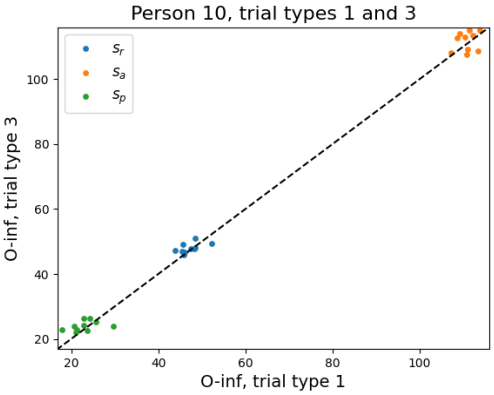


Figure D.31: O-information of all repeats of trial type 1 and 3 for person 10.

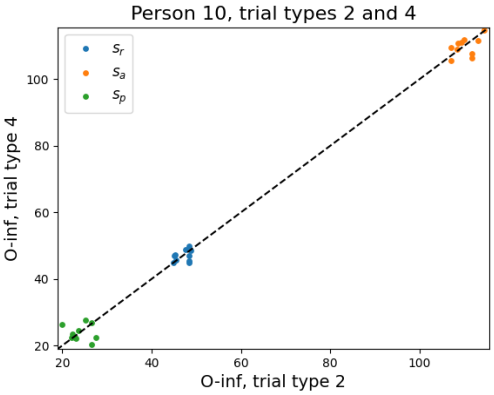


Figure D.32: O-information of all repeats of trial type 2 and 4 for person 10.

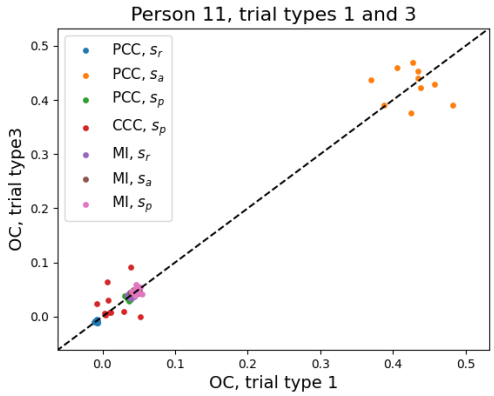


Figure D.33: OCs of all repeats of trials 1 and 3 for person 11.

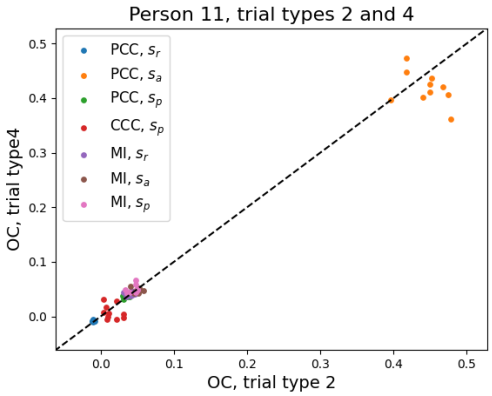


Figure D.34: OCs of all repeats of trial type 2 and 4 for person 11.

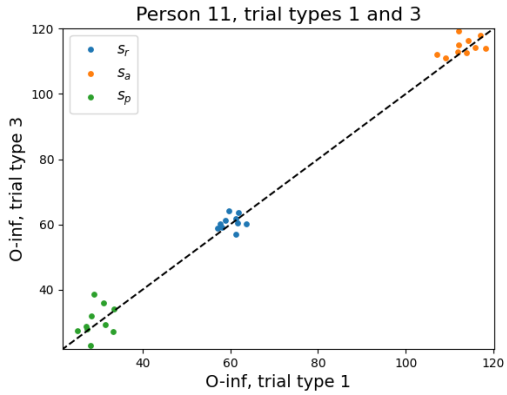


Figure D.35: O-information of all repeats of trial type 1 and 3 for person 11.

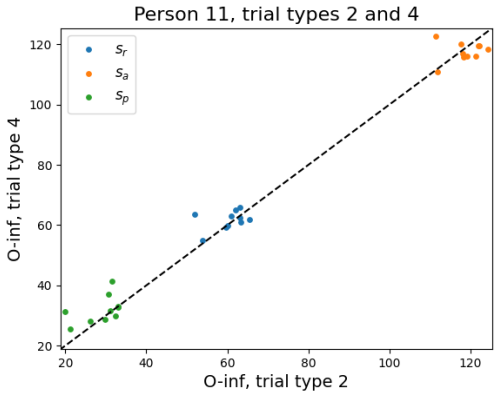


Figure D.36: O-information of all repeats of trial type 2 and 4 for person 11.

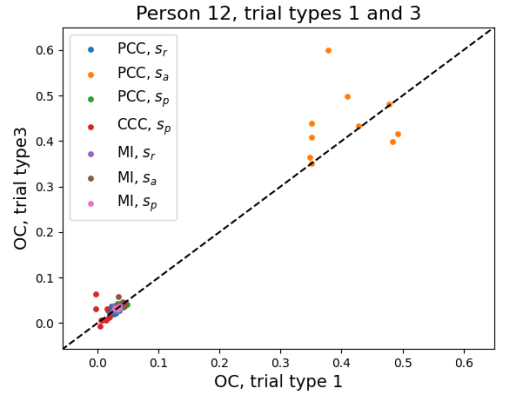


Figure D.37: OCs of all repeats of trials 1 and 3 for person 12.

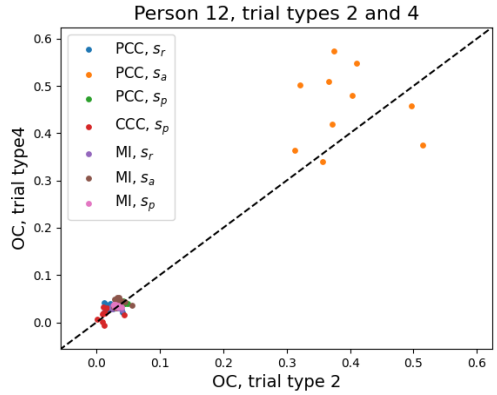


Figure D.38: OCs of all repeats of trial type 2 and 4 for person 12.

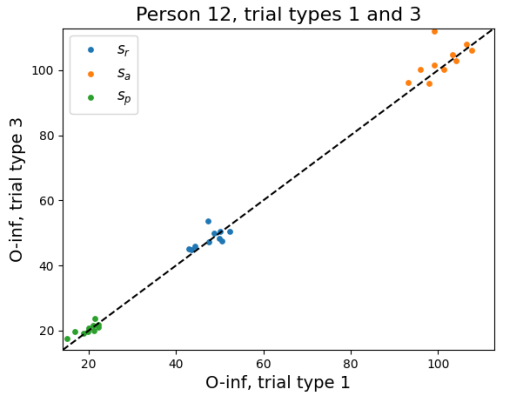


Figure D.39: O-information of all repeats of trial type 1 and 3 for person 12.

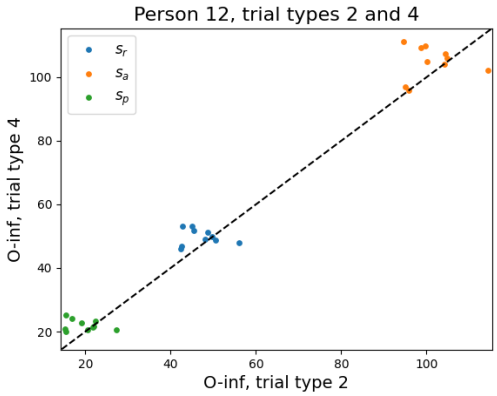


Figure D.40: O-information of all repeats of trial type 2 and 4 for person 12.

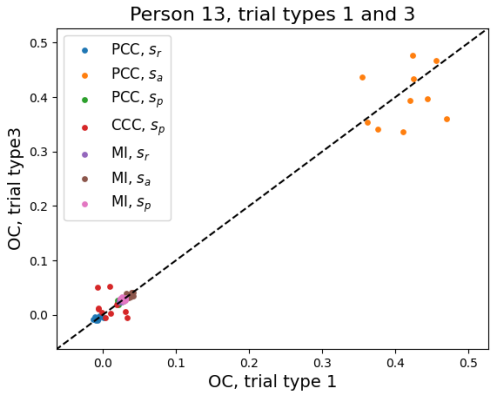


Figure D.41: OCs of all repeats of trials 1 and 3 for person 13.

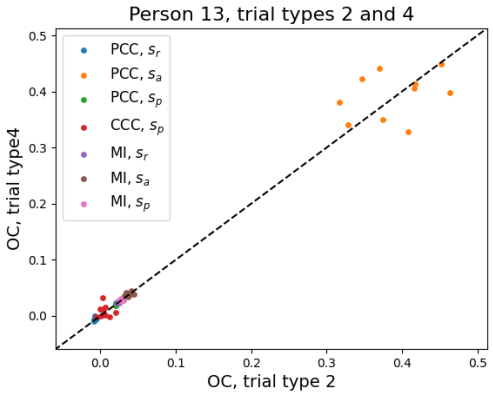


Figure D.42: OCs of all repeats of trial type 2 and 4 for person 13.

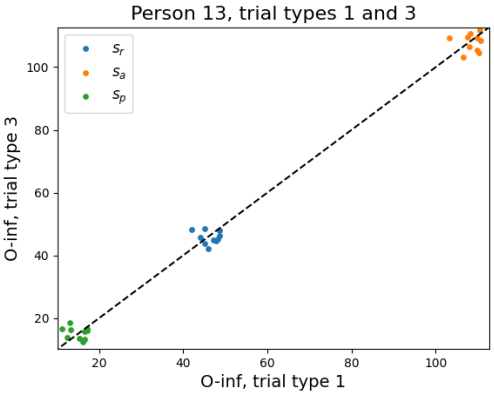


Figure D.43: O-information of all repeats of trial type 1 and 3 for person 13.

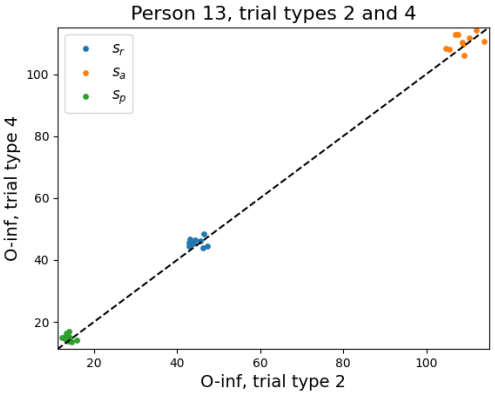


Figure D.44: O-information of all repeats of trial type 2 and 4 for person 13.

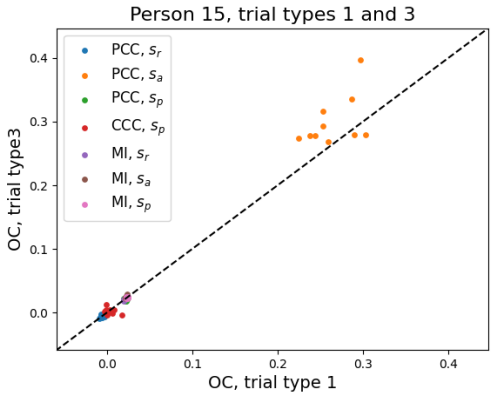


Figure D.45: OCs of all repeats of trials 1 and 3 for person 15.

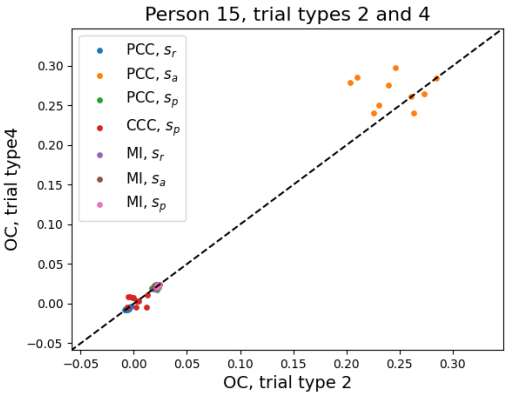


Figure D.46: OCs of all repeats of trial type 2 and 4 for person 15.

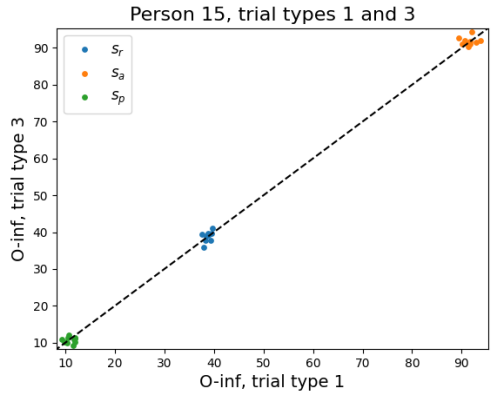


Figure D.47: O-information of all repeats of trial type 1 and 3 for person 15.

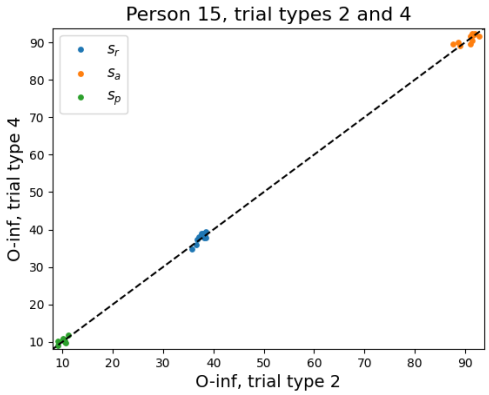


Figure D.48: O-information of all repeats of trial type 2 and 4 for person 15.

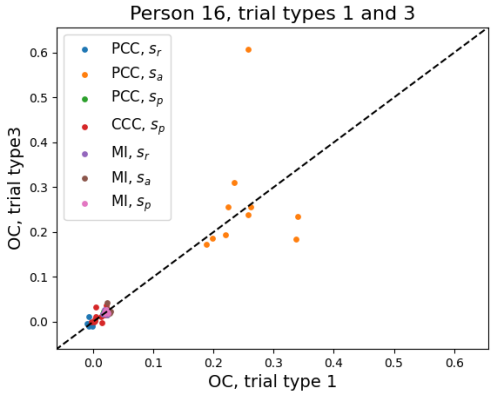


Figure D.49: OCs of all repeats of trials 1 and 3 for person 16.

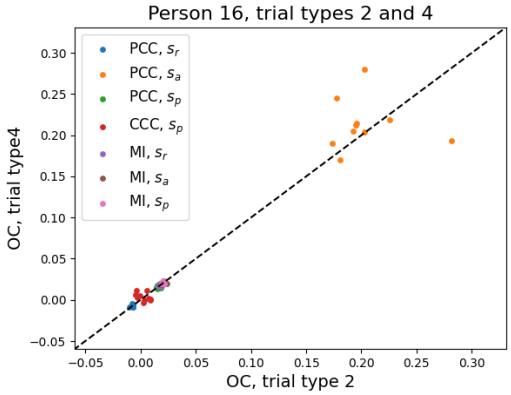


Figure D.50: OCs of all repeats of trial type 2 and 4 for person 16.

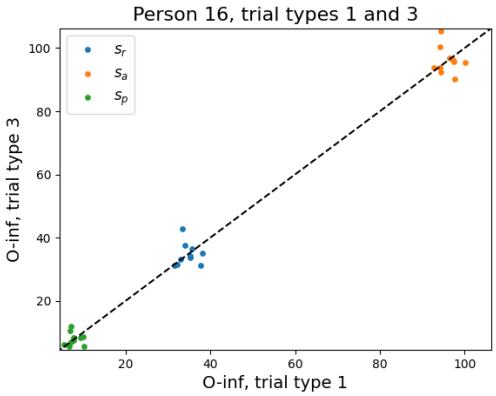


Figure D.51: O-information of all repeats of trial type 1 and 3 for person 16.

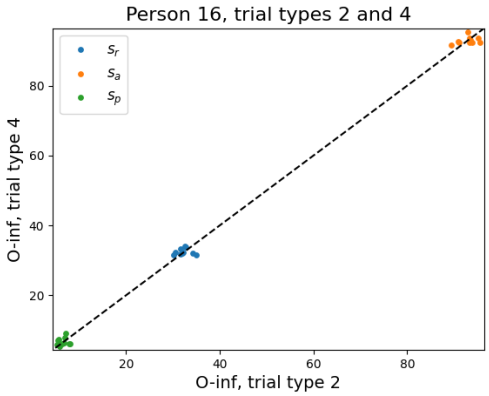


Figure D.52: O-information of all repeats of trial type 2 and 4 for person 16.

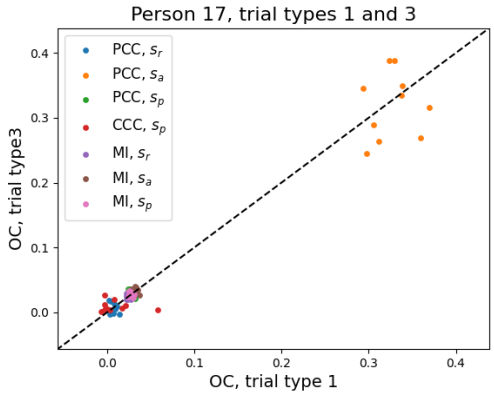


Figure D.53: OCs of all repeats of trials 1 and 3 for person 17.

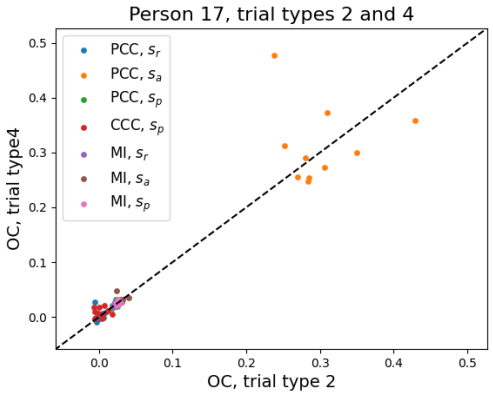


Figure D.54: OCs of all repeats of trial type 2 and 4 for person 17.

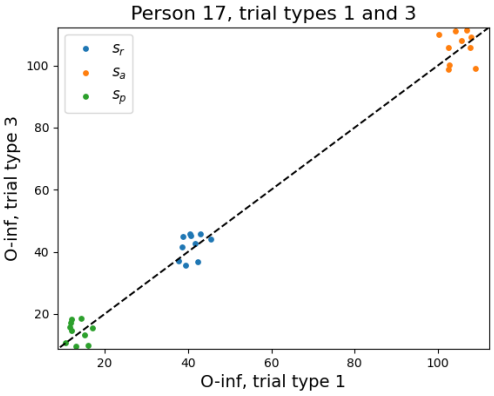


Figure D.55: O-information of all repeats of trial type 1 and 3 for person 17.

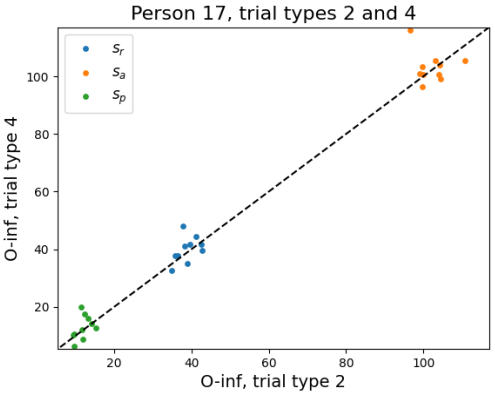


Figure D.56: O-information of all repeats of trial type 2 and 4 for person 17.

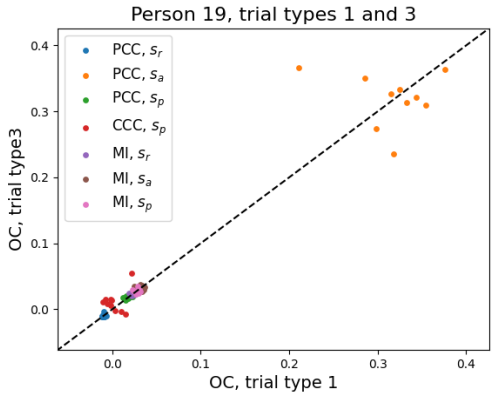


Figure D.57: OCs of all repeats of trials 1 and 3 for person 19.

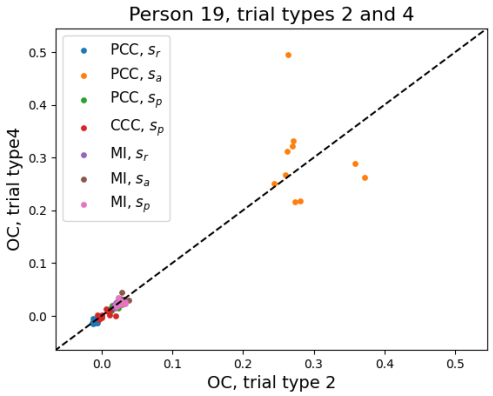


Figure D.58: OCs of all repeats of trial type 2 and 4 for person 19.

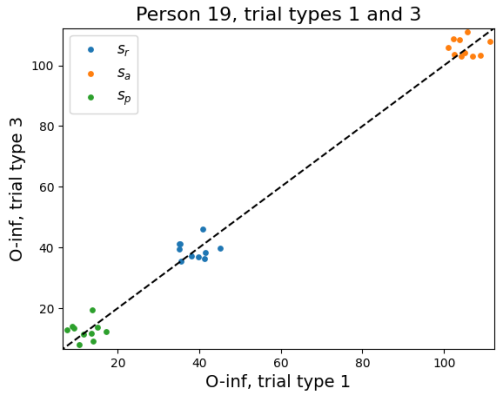


Figure D.59: O-information of all repeats of trial type 1 and 3 for person 19.

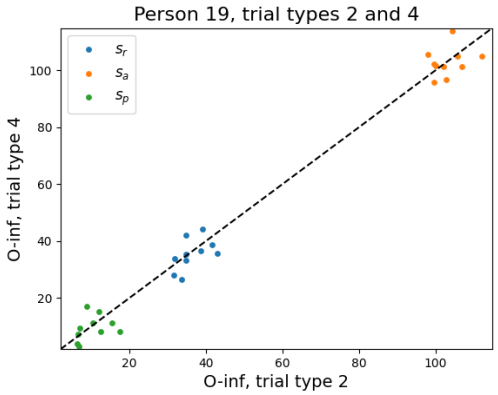


Figure D.60: O-information of all repeats of trial type 2 and 4 for person 19.

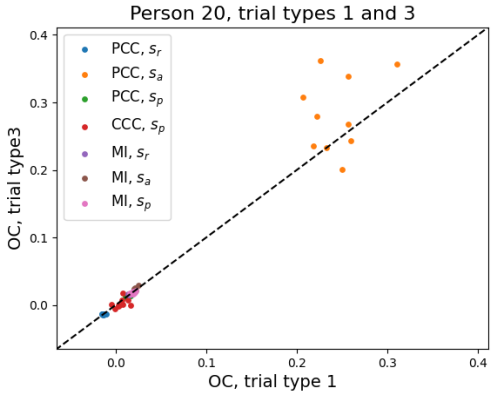


Figure D.61: OCs of all repeats of trials 1 and 3 for person 20.

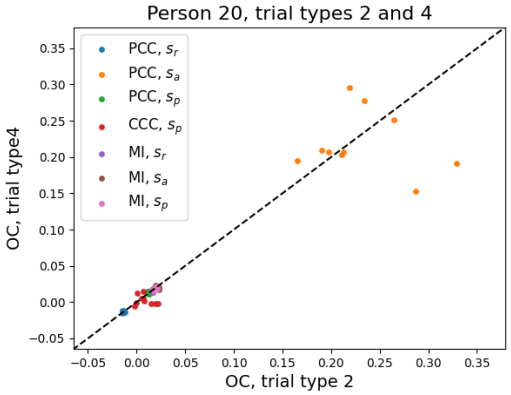


Figure D.62: OCs of all repeats of trial type 2 and 4 for person 20.

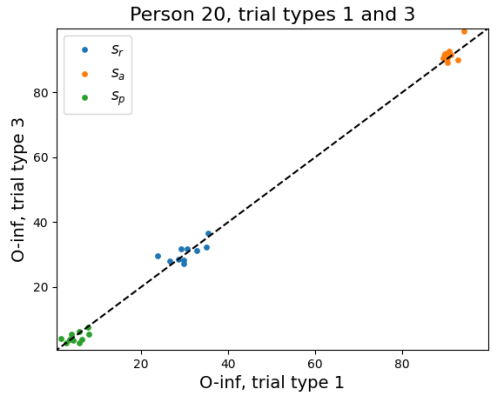


Figure D.63: O-information of all repeats of trial type 1 and 3 for person 20.

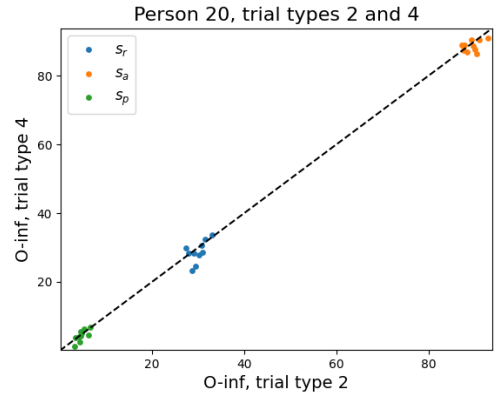


Figure D.64: O-information of all repeats of trial type 2 and 4 for person 20.

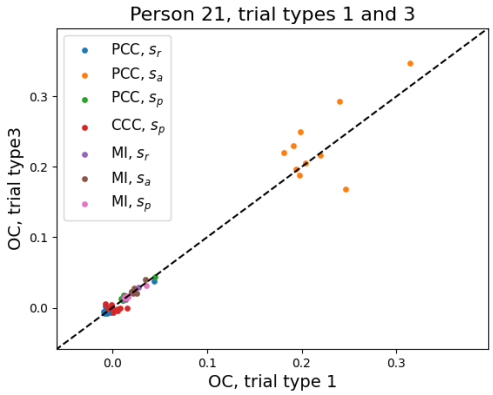


Figure D.65: OCs of all repeats of trials 1 and 3 for person 21.

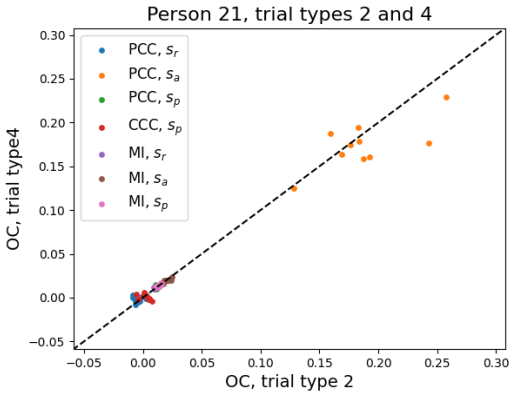


Figure D.66: OCs of all repeats of trial type 2 and 4 for person 21.

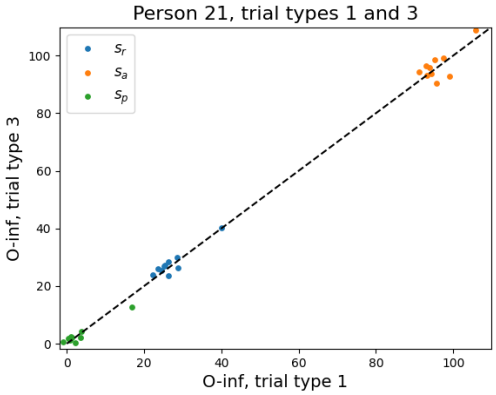


Figure D.67: O-information of all repeats of trial type 1 and 3 for person 21.

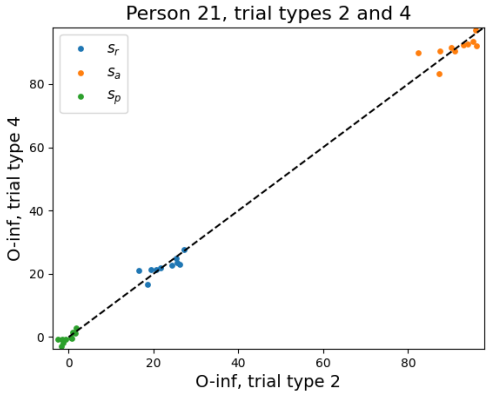


Figure D.68: O-information of all repeats of trial type 2 and 4 for person 21.

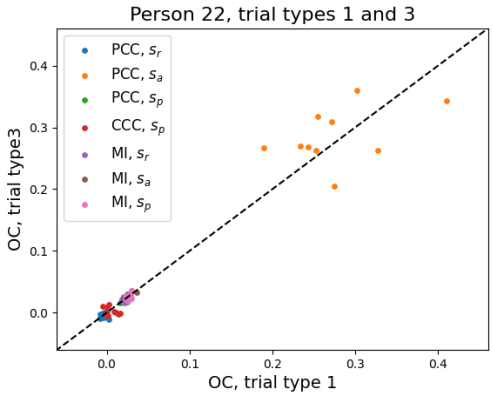


Figure D.69: OCs of all repeats of trials 1 and 3 for person 22.

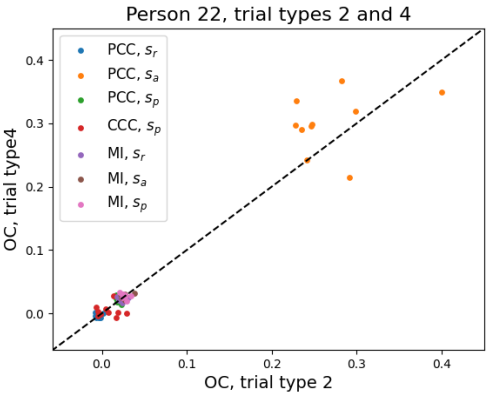


Figure D.70: OCs of all repeats of trial type 2 and 4 for person 22.

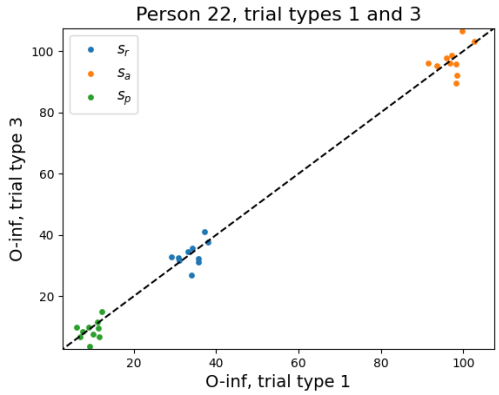


Figure D.71: O-information of all repeats of trial type 1 and 3 for person 22.

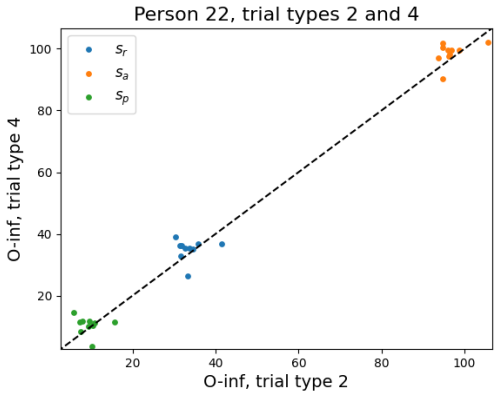


Figure D.72: O-information of all repeats of trial type 2 and 4 for person 22.