Estimation of Transfer Entropy

Giannarakis, G.

Award date:
2020

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science

Estimation of Transfer Entropy

Master Thesis

Georgios Giannarakis

Supervisors:
Alessandro Di Bucchianico (TU/e)
Errol Zalmijn (ASML)
Hans Onvlee (ASML)

Eindhoven, July 2020

Abstract

ASML is the world's leading provider of lithography systems for the semiconductor industry, manufacturing complex machines that are critical to the production of integrated circuits or chips. Miniaturization of microprocessors is connected to growing complexity of lithography systems, imposing greater challenges on their design, prognostics and diagnostics. The ASML Research department plays an important role in investigating original concepts and applications that drive technological breakthroughs. Current ASML research focuses on probing data-driven techniques that unravel the structure of the internal dynamics of lithography systems, in order to gain novel insights into system behavior. Applying causal inference techniques such as the information-theoretic transfer entropy on time series data generated by lithography systems is a promising way to deal with this task.

This thesis investigates transfer entropy in the case of non-stationary time series, including a theoretical analysis as well as practical estimation and insights. A concrete mathematical system is studied, satisfying the additional condition of increment stationarity. Exact results are derived, succeeded by the examination of a transfer entropy estimator in this system. Provided a moderate amount of data are available, the estimator closely approximates the theoretical transfer entropy values, however bias correction is required. This thesis also develops and executes a benchmark study for the objective comparison of different causal inference methods in time series data. Methods are qualitatively classified based on a set of important properties compiled, while their performance and running times are carefully evaluated. Methods assessed generally exhibit satisfactory performance, with the information-theoretic techniques tested ranking at the top.

Keywords: information theory, transfer entropy, causal inference, Granger causality, time series, stationarity, random walk, dynamical systems

Acknowledgements

Concluding this six month graduation project at ASML, I would like to express my gratitude to many people. First, to my academic advisor Alessandro Di Bucchianico, whose expert guidance, support and clear communication immensely benefited the project, properly orienting and keeping me on the right track.

I would also like to thank my ASML supervisor Errol Zalmijn. The passion for the project he always showed and the encouragement he graciously provided me were deeply inspirational. Then, I thank Hans Onvlee, for giving me the opportunity to pursue this project at ASML, and Pierluigi Frisco, for meticulously informing me about the miscellaneous details a student project at ASML entails.

I also want to thank my parents Nikos and Froso and my sister Eirini, for their perpetual support and love. My studies wouldn't have been the same without Eleni, whom I thank for her understanding and patience.

Georgios Giannarakis
Athens, July 2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to ASML

ASML is the world's leading provider of lithography systems for the semiconductor industry, manufacturing complex machines that are critical to the production of integrated circuits or chips. ASML was founded in 1984 by Philips and Advanced Semiconductor Materials International with the aim of developing lithography systems for the growing semiconductor market. It is a multinational company with over 60 locations in 16 countries worldwide, on aggregate employing more than 24,000 people.

Technology

A lithography system uses light to print tiny patterns on silicon, an essential step in the mass production of computer chips. Light is projected through a blueprint of the pattern to be printed, and is subsequently focused onto a photosensitive silicon wafer. After the pattern is printed, the wafer is slightly moved and another copy is made. This process is repeated to fully cover the wafer in patterns, comprising one layer of the wafer's chips.

The wavelength of the light used dictates the type of the lithography system: ASML is the world's sole manufacturer of lithography systems employing Extreme Ultraviolet Light (EUV) with a wavelength of 13.5 nanometers (comparable to that of an X-ray) offering significant improvements compared to the older Deep Ultraviolet (DUV) lithography systems that are also in ASML production. A DUV system is shown in Figure 1.1.

Essential to ASML technology are the metrology solutions developed to rapidly measure imaging performance on wafers. Relevant data are fed back to the system in real-time, safeguarding the production performance of lithography systems. Inspection tools help in detecting and analyzing defects that are located among billions of printed patterns. Moreover, ASML develops pioneering software that aids the manufacturing process - elevating lithography systems from high-tech hardware to a hybrid of high-tech hardware and advanced software.

Research at ASML

Lithography systems are among the most complex systems manufactured today and ASML's Research department plays an important role in investigating novel concepts and applications to drive technological breakthroughs. Within the Technology corporate function, the Research department aims at creating, developing and demonstrating technology solutions that further explore, extend and improve existing ASML technology roadmaps. After providing proof of concept, Research results are transferred to other ASML corporate functions such as Development and Engineering or System Engineering. This Master's thesis was conducted within the Software & Data Science team of the Research department.

Figure 1.1: The latest DUV lithography system NXT:2000i as seen on the website of ASML

1.2 Project description

Miniaturization of microprocessors goes hand in hand with growing complexity of lithography systems. The underlying physical mechanisms become increasingly complicated to fully understand, imposing greater challenges on the design, prognosis and diagnosis of such systems.

Lithography systems are characterized by highly nonlinear dynamics observed over a large parameter space across multiple time scales. Critical requirements include position control with nanometer precision and temperature control with milli-Kelvin accuracy even during rapid acceleration of system modules.

Correlation studies as well as model-based approaches may prove inadequate to capture nonlinear causal dependencies in such complex systems, as correlation cannot prove causation and prior model assumptions are often invalid. Model-free approaches such as the Information Theory based *Transfer Entropy* do not rely on assumptions regarding underlying physical mechanisms, but inevitably come with high computational costs.

In current ASML research, transfer entropy is used to identify causal interactions between time series from ASML (sub-)systems, in order to gain better understanding of the system's physical behavior. New insights are key to enable reliable diagnostics and predictive maintenance or overall system performance optimization through effective design improvements.

Investigating the hypothesis that transfer entropy is a viable measure of causality in lithography systems, this thesis addresses the following research questions:

- ASML lithography system time series are often non-stationary, featuring e.g. drifts or degrading processes. However, current transfer entropy estimators typically assume stationarity of input data. How to estimate transfer entropy in non-stationary time series?

- A wide range of causal inference methods has been proposed over the years, coming from a diverse group of mathematical theories. Each of these methods has its own merits and limitations considering different criteria of causal inference. Which characteristics can be used to classify causal inference methods, and which criteria can be used to compare and contrast their performance? Although transfer entropy has demonstrated to be a promising measure of causality in ASML lithography systems, it is important to benchmark its performance against other causal inference methods. How to develop and what are the results of a benchmark study featuring several causal inference methods, including transfer entropy?

1.3 Report outline

Succeeding the current chapter, a theoretical introduction featuring the mathematical background and details relevant to the project comprises Chapter 2. Chapter 3 contains a discussion of estimators in Information Theory, thereby setting up a comprehensive study of transfer entropy in a non-stationary setting presented in Chapter 4 that pertains to the first research question. The report subsequently shifts to the second research question, commencing with a discussion of data in Chapter 5. A benchmark framework corresponding to the second research question is developed in Chapter 6 and its results are presented and discussed in Chapter 7. The report finishes with Chapter 8 where conclusions are drawn, results are summarized and further research questions are formulated.

Chapter 2

Theoretical Background

This chapter contains a comprehensive discussion of the relevant mathematical theory that is used in this project. It introduces the mathematical fields of information theory and causal inference as well as specific topics in the field of stochastic processes and time series. Supplementary knowledge relating to the contents of this chapter is given in Appendix A.

2.1 Information theory

The field of information theory was pioneered by C. Shannon in his landmark article Shannon (1948), where a mathematical treatment of communication was presented and relevant terms such as the entropy of a random variable were introduced. The following are based on Cover and Thomas (2006), one of the main references for information theory, as well as Bossomaier et al. (2016).

2.1.1 Shannon entropy

Consider a discrete random variable X and its image $\mathcal{X}$ that contains its (countable) values. Let $p_X(x) = P(X = x)$ be the probability mass function of X. The *information content* of an $x \in \mathcal{X}$ is defined as

$$h(x) = -\log p_X(x) \tag{2.1}$$

The entropy of a random variable is then the average information content of the variable, and it can be thought of as the average information or uncertainty of this random variable. Formally,

Definition 2.1.1 (Shannon Entropy)**.** *The Shannon entropy of a discrete random variable X with a probability mass function p_X is defined as*

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \tag{2.2}$$

The selection of the logarithmic function in defining the above can be rigorously derived starting from a general entropy form and stipulating an axiom (see Appendix A).

In the following, the subscript in p_X may be omitted given that the variable we refer to is clear. When a logarithm with base 2 is used, entropy is measured in *bits*. In his original formulation, Shannon used natural logarithms. In that case, entropy is measured in *nats*. Throughout the report, log denotes the natural logarithm, and other bases are explicitly denoted with a subscript.

Example 2.1.2. *Consider a random variable following a discrete uniform distribution over 32 outcomes, i.e. $X \sim \mathcal{U}(\{1, 32\})$. The entropy of this random variable is*

$$H(X) = -\sum_{i=1}^{32} p(i) \log_2 p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log_2 \frac{1}{32} = \log_2 32 = 5 \text{ bits.} \tag{2.3}$$

Shannon's original article pertains to the mathematical formalization of communication. Within this context, entropy is defined as a means to studying the communication of a source with a destination through a channel. While the field of signal processing that intertwines Shannon's theory of communication is out of scope for this project, a short remark is now given on interpreting the above example from the perspective of data compression.

Intuitively, to be able to identify an outcome of this variable, a label that can take 32 different values is needed. A five-dimensional binary vector (that is, a 5-bit string) is therefore enough, as it can be used to encode $2^5 = 32$ different values. This is not coincidental; there is a deep connection between the entropy of a random variable and the length of codes that are able to describe them (Cover and Thomas, 2006, Chapter 5).

Entropy can be naturally extended to two (or more) random variables by simply considering them as a single random vector.

Definition 2.1.3 (Joint Entropy)**.** *The joint entropy of two discrete random variables X and Y is*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{2.4}$$

Joint entropy is measuring the uncertainty included in the random vector (X,Y).

A key quantity to define is *conditional entropy*: the uncertainty left in a random variable after we have taken into account some context.

Following the idea of the definition of conditional expectation, first the *conditional entropy of X given that $Y = y$* is defined. This is done by utilizing the conditional probability mass function $p(x|y)$:

$$H(X|y) = -\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \tag{2.5}$$

Note that $H(X|y)$ is a function of y. To get the conditional entropy of X given Y we then simply average over y:

Definition 2.1.4 (Conditional Entropy)**.** *The conditional entropy of X given Y, where X and Y are discrete random variables is given by:*

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \tag{2.6}$$

A useful result that connects joint and conditional entropy is the following chain rule:

Theorem 2.1.5 (Chain Rule)**.** *For two discrete random variables the following holds:*

$$H(X,Y) = H(X) + H(Y|X) \tag{2.7}$$

Shannon entropy can be extended to the case of continuous random variables. In that case, it is known as *differential entropy*.

Theorem 2.1.6 (Differential Entropy)**.** *The differential entropy $h(X)$ of a continuous random variable X with probability density function f is defined as*

$$h(X) = -\int_A f(x)\log f(x)dx \tag{2.8}$$

where A is the support of the density f of X, namely $A = \{x \in \mathcal{X} : f(x) > 0\}$

Note that the integral need not necessarily exist, and contrary to the discrete case, it can be negative.

Example 2.1.7. *As an example, the differential entropy of a normally distributed random variable is calculated below: Let* $X \sim \mathcal{N}(0, \sigma^2)$*. The density of this random variable is:*

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{x^2}{2\sigma^2}\big) \quad , x \in \mathbb{R} \tag{2.9}$$

Then,

$$\begin{aligned} h(X) &= -\int_{\mathbb{R}} \varphi(x) \log \varphi(x) dx \\ &= -\int_{\mathbb{R}} \varphi(x) \left[-\frac{x^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2}) \right] dx \\ &= \frac{E[X^2]}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \\ &= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2) \\ &= \frac{1}{2}\log e + \frac{1}{2}\log(2\pi\sigma^2) \\ &= \frac{1}{2}\log(2\pi e\sigma^2) \quad \textit{nats.} \end{aligned} \tag{2.10}$$

To derive the differential entropy in bits, the base of the logarithm is changed from e *to* 2*:*

$$h(X) = \frac{1}{2}\log_2 2\pi e\sigma^2 \quad \textit{bits.} \tag{2.11}$$

Just like its discrete counterpart H, differential entropy h can be extended to *joint* and *conditional* differential entropy in a similar way. The chain rule for differential entropy also exists. The same holds for *mutual information*, a discussion of which from the discrete perspective follows.

Note that for any given dataset, the calculation of entropy and other relevant information theoretic quantities simply involves the estimation of probability functions. Therefore, when information theory techniques are employed for the study of a dataset, no concrete assumptions about the relations between the variables in the form of a model are needed. In that sense information theory methods are model-free.

At the same time, the absence of model assumptions combined with a potential high-dimensionality of information-theoretic quantities imposes significant difficulty to their estimation; this is the subject of Chapter 3.

2.1.2 Mutual information

Intuitively, $H(X)$ is the uncertainty in X, while $H(X|Y)$ is the uncertainty that remains in X after observing Y. It is also sensible to be interested in the reduction of uncertainty in X due to the knowledge of Y.

This is exactly the notion of *mutual information*: the amount of information that is *shared* between two random variables X and Y. Mutual information is a measure of their statistical dependence, a generalized version of the correlation coefficient to the non-linear case.

Definition 2.1.8 (Mutual Information)**.** *The mutual information of two discrete random variables* X *and* Y*, is given by:*

$$I(X;Y) = H(X) - H(X|Y) \tag{2.12}$$

Taking into account (2.7), it is easily seen that $H(X) - H(X|Y) = H(Y) - H(Y|X)$, which makes mutual information symmetric in X and Y.

Expanding the above definition by substituting the analytical formulas for entropy and conditional entropy, mutual information admits a convenient form, that can also be expressed via the *Kullback-Leibler divergence* measure:

Definition 2.1.9 (K-L Divergence)**.** *Given two discrete random variables defined on the same probability space with respective probability mass functions p and q. If $q(x) = 0$ implies $p(x) = 0$ $\forall x \in \mathcal{X}$, then the Kullback-Leibler (K-L) divergence is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{2.13}$$

with the convention $0 \log \frac{0}{0} = 0$ and $D(p||q) = +\infty$ if $\exists\ x \in \mathcal{X} : q(x) = 0$ and $p(x) > 0$.

The Kullback-Leibler divergence is not symmetric, nor does it satisfy the triangle inequality. However, it can be loosely thought of as the *distance* between the probability distributions p and q. This is also encouraged by the following result that we prove in Appendix A:

Theorem 2.1.10. *Let p and q be two probability mass functions defined on the same probability space. Then*

$$D(p||q) \geq 0 \tag{2.14}$$

with equality if and only if $p(x) = q(x)$ for all x.

Now, for the random variables X and Y, mutual information is the distance (in the Kullback-Leibler sense) of the joint probability function $p_{(X,Y)}(x, y)$ from the product of the marginal probability functions $p_X(x)p_Y(y)$ which we denote with $p_X \times p_Y$ in the K-L operator.

$$I(X;Y) = D(p_{(X,Y)}||p_X \times p_Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{(X,Y)}(x, y) \log \frac{p_{(X,Y)}(x, y)}{p_X(x)p_Y(y)} \tag{2.15}$$

The above results yield a characterization of independence through mutual information. Indeed, for random variables X and Y, using the expression (2.15) and theorem (2.1.10) we infer that $I(X;Y) = 0$ if and only if $p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$, that is, if and only if, X and Y are independent. The following corollary is thus proven:

Corollary 2.1.11. *The random variables X and Y are independent if and only if $I(X;Y) = 0$*

In that sense, mutual information quantifies the distance of X and Y from independence, justifying its interpretation as a measure of dependence. Another interesting corollary follows from Theorem 2.1.10. In Example 2.1.2 we calculated the Shannon entropy for a discrete uniform random variable X with an image $\mathcal{X}$. Its Shannon entropy was found to be equal to 5 which is equal to $\log_2 32$, while 32 was the cardinality of $\mathcal{X}$. This was not a coincidence; we will now prove that this value was the maximum possible entropy for a discrete probability distribution defined over $\mathcal{X}$.

Corollary 2.1.12. *Let X be a discrete random variable, and $\mathcal{X}$ be its image with a finite cardinality $|\mathcal{X}|$. Then, $H(X) \leq \log\big(|\mathcal{X}|\big)$, with equality if and only if X has the discrete uniform distribution over $\mathcal{X}$.*

Proof. Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the probability mass function of the discrete uniform distribution over $\mathcal{X}$, and let p be an arbitrary probability mass function of X. We write:

$$D(p||u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log\big(\frac{1}{|\mathcal{X}|}\big) = \log\big(|\mathcal{X}|\big) - H(X) \tag{2.16}$$

From Theorem 2.1.10 we get that $\log\big(|\mathcal{X}|\big) - H(X) \geq 0$. The result follows by observing that $\log\big(|\mathcal{X}|\big)$ is the entropy of the discrete uniform distribution over $\mathcal{X}$. This can be easily proven through a direct calculation such as the one featured in example 2.1.2. □

Since mutual information is directly defined through (conditional) entropy, extending mutual information to *conditional mutual information* is straightforward:

Definition 2.1.13 (Conditional Mutual Information)**.** *Let X, Y, Z be discrete random variables. The conditional mutual information of X and Y given Z is*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{2.17}$$

The conditional mutual information of the random variables X and Y conditioned on Z is the information that is shared between X and Y *in the context* of Z.

If mutual information being zero characterized the independence of X and Y, conditional mutual information being zero characterizes the *conditional independence of X and Y given Z.*

2.1.3 Transfer entropy

Mutual information quantifies the information that is shared between two static random variables. However, in applications as well as in research, it is very often the case where *time-dynamic* processes are considered, and data from multiple sources are registered over time.

The extension of the idea behind mutual information to the time-dynamic case, was conceptualized within the context of information theory as the quantification of the *information transfer* between different time series.

Attempting to formalize a measure for the transfer of information from a time series Y_t (the *source*) to a time series X_t (the *target*), T. Schreiber proposed the notion of *transfer entropy* in Schreiber (2000).

Throughout the report, transfer entropy (TE) is considered in discrete time. This is also the case for the overwhelming majority of literature. Recent advances on continuous time transfer entropy exist (Spinney et al. (2017), Cooper and Edgar (2019)) but they are out of scope for this project.

To define TE following the original formulation of Schreiber, first a Markovian assumption has to be made. We thus define:

Definition 2.1.14 (Markov chain of order m)**.** *A discrete time stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is a Markov chain of order m when, for any $t > m$, the following property holds:*

$$P(X_t = x_t|X_{t-1} = x_{t-1}, ..., X_1 = x_1) = P(X_t = x_t|X_{t-1} = x_{t-1}, ..., X_{t-m} = x_{t-m}) \tag{2.18}$$

That is, the future of such a process only depends on its past m states. As noted above, TE will always be considered in discrete time; in the following, the terms Markov *process* and Markov *chain* are therefore used interchangeably.

To define TE, it is assumed that the source Y_t is a Markov process of order ℓ, and the target X_t is a Markov process of order k. Therefore, the future state of the source and target only depends on their past ℓ and k states respectively. Note that X_t is allowed to depend on the future of Y_t and information might still be getting transferred from Y_t to X_t; in fact, this is what TE aspires to investigate.

Remark. *Notice here the implicit assumption that the future value of the target X_t depends only on its past states or on both its past states and the past states of the source Y_t - there is no third process Z_t interfering with the target Gencaga et al. (2015). This constraint is removed with the introduction of conditional transfer entropy.*

Before proceeding with transfer entropy the notion of *embedding vectors* is first defined.

Definition 2.1.15 (Embedding Vector)**.** *Let $\{U_t\}_{t\in\mathbb{Z}}$ be a time series. The embedding vector $U_t^{(d,\tau)}$ is the following random vector of past states of U_t:*

$$U_t^{(d,\tau)} = (U_t, U_{t-\tau}, U_{t-2\tau}, ..., U_{t-(d-1)\tau}) \tag{2.19}$$

The embedding vector notation $U_t^{(d,\tau)}$ can be simplified to $U_t^{(d)}$ when $\tau = 1$, which yields the embedding vector $(U_t, U_{t-1}, U_{t-2}, ..., U_{t-(d-1)})$. In literature, the parameter d is called the *embedding dimension* and τ is called the *embedding delay.*

Now, transfer entropy can be defined.

Definition 2.1.16 (Transfer Entropy)**.** *At time t, the transfer entropy from the ℓ^{th} order Markov process Y_t (the source) to the k^{th} order Markov process X_t (the target) is defined as follows:*

$$T_{Y\to X}^{(k,\ell)}(t) = I(X_t; Y_{t-1}^{(\ell)} | X_{t-1}^{(k)}) \tag{2.20}$$

Note that in (2.20), k and ℓ are both embedding dimensions and are still denoted with the superscript (k,ℓ) since the embedding delay $\tau = 1$ and is therefore omitted. Furthermore, for stationary processes (see Section 2.3) the time index t can be omitted.

Remark. *The introduction of embedding vectors given the Markovian assumption of (2.20) may appear as mere notational convenience. Indeed, the Markovian context formulated here is naturally associated with the embedding vectors $Y_{t-1}^{(\ell)}$, $X_{t-1}^{(k)}$ since they capture the memory of each process. However, for the general case and in real data where a similar Markovian assumption might be invalid, the discussion of embedding vectors is much deeper - and interconnected with the theory of dynamical systems Takens (1981), Kantz and Schreiber (2006). We therefore note that the Markovian assumption that is made here mostly serves simplification purposes - all definitions and results of this section still hold without it.*

According to the mutual information interpretation discussed before, transfer entropy is *the information that is shared between the current state of the target and the past states of the source, in the context of the target's own past.* Note that TE is not symmetric in X and Y. Thus, it is appropriate for capturing the *directed information transfer* between two processes. This notion of directionality is also of paramount importance to the causal interpretation of TE that follows.

Transfer entropy is therefore a form of conditional mutual information. Using (2.7) and (2.17), it can be simplified to a combination of joint and marginal entropies:

$$\begin{aligned} T_{Y\to X}^{(k,\ell)}(t) &= I(X_t; Y_{t-1}^{(\ell)} | X_{t-1}^{(k)}) \\ &= H(X_t | X_{t-1}^{(k)}) - H(X_t | X_{t-1}^{(k)}, Y_{t-1}^{(\ell)}) \\ &= H(X_t, X_{t-1}^{(k)}) - H(X_{t-1}^{(k)}) - H(X_t, X_{t-1}^{(k)}, Y_{t-1}^{(\ell)}) + H(X_{t-1}^{(k)}, Y_{t-1}^{(\ell)}) \end{aligned} \tag{2.21}$$

Besides the interpretation of TE stemming from the conditional mutual information definition (2.20) given above, the second equality of (2.21) provides another interpretation of TE in terms of conditional entropy. Recall that conditional entropy $H(X|Y)$ is the uncertainty left in X after accounting for Y, or in other words, the degree of uncertainty of X *resolved* by Y. Therefore, TE may equivalently be understood as *the degree of uncertainty of X resolved by the past of Y over and above the degree of uncertainty of X resolved by its own past.*

Since TE is a form of conditional mutual information, *conditioning* on a third process $Z = Z_t$ when examining the information transfer $Y \to X$ from source Y to target X is trivially done by simply adding Z in the conditional part of (2.20).

This enables the definition of *conditional transfer entropy.*

Definition 2.1.17 (Conditional transfer entropy)**.** *At time t, the conditional transfer entropy from the ℓ^{th} order Markov process Y_t to the k^{th} order Markov process X_t given the m^{th} order Markov process Z_t is defined as:*

$$T_{Y\to X|Z}^{(k,\ell,m)}(t) = I(X_t; Y_{t-1}^{(\ell)} | X_{t-1}^{(k)}, Z_{t-1}^{(m)}) \tag{2.22}$$

In his original formulation, Schreiber gives an equivalent analytic definition for TE which we prove in Appendix A as a theorem. In the following, the letter p is used to denote different probability mass functions, to avoid overloading notation. For example, $p(x_{t-1}^{(k)}) = p_{X_{t-1}^{(k)}}(x_{t-1}^{(k)})$, while $p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) = p_{(X_t, X_{t-1}^{(k)}, Y_{t-1}^{(\ell)})}(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)})$

Theorem 2.1.18 (Transfer Entropy - Analytic)**.** *As defined in (2.20), transfer entropy admits the following analytic form:*

$$T_{Y\to X}^{(k,\ell)}(t) = \sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(l)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log \frac{p(x_t | x_{t-1}^{(k)}, y_{t-1}^{(\ell)})}{p(x_t | x_{t-1}^{(k)})} \tag{2.23}$$

To better interpret the analytic form of transfer entropy, we can decompose (2.23) into:

$$T^{(k,\ell)}_{Y\to X}(t) = \sum_{x^{(k)}_{t-1}, y^{(\ell)}_{t-1}} p(x^{(k)}_{t-1}, y^{(\ell)}_{t-1}) \sum_{x_t} p(x_t|x^{(k)}_{t-1}, y^{(\ell)}_{t-1}) \log \frac{p(x_t|x^{(k)}_{t-1}, y^{(\ell)}_{t-1})}{p(x_t|x^{(k)}_{t-1})} \tag{2.24}$$

Note that the inner sum is the K-L divergence between the distributions $X_t|(X^{(k)}_{t-1}, Y^{(l)}_{t-1})$ and $X_t|X^{(k)}_{t-1}$, i.e. the deviation of the target X_t from independence from (the past of) a source Y_t in the context of the target's own past. Then, TE is this K-L divergence averaged over the distribution of the past states $(X^{(k)}_{t-1}, Y^{(l)}_{t-1})$.

Recalling that any K-L divergence is non-negative (Theorem 2.1.10) makes transfer entropy a non-negative measure of directed information transfer. Moreover, combining the fact that conditional mutual information characterizes the notion of conditional independence (see comments below (2.17)), and the analytic form of TE (2.23), it can be seen that TE also characterizes a specific conditional independence relation between the source and the target:

$$T^{(k,l)}_{Y\to X}(t) = 0 \iff \tag{2.25}$$

$$I(X_t; Y^{(\ell)}_{t-1}|X^{(k)}_{t-1}) = 0 \iff \tag{2.26}$$

$$p(x_t|x^{(k)}_{t-1}, y^{(l)}_{t-1}) = p(x_t|x^{(k)}_{t-1}) \iff \tag{2.27}$$

$$(x_t \perp\!\!\!\perp y^{(\ell)}_{t-1}) \mid x^{(k)}_{t-1} \tag{2.28}$$

That is, the transfer entropy from source Y to target X being zero is equivalent with the present of the target being independent of the source's past *in the context of* the target's own past.

Since its introduction, TE has attracted significant attention of both practitioners and researchers in a variety of scientific fields ranging from neuroscience to finance and engineering (e.g. Vicente et al. (2010), Papana et al. (2015), Bauer et al. (2007)).

The prominence of TE is largely due to a very specific quality it carries as a measure of directed information transfer: a causal interpretation. Transfer entropy therefore establishes a connection between Information Theory and Causal Inference. This statement and the concepts involved are elaborated in the following section, and a succinct presentation of the causal inference theory that is relevant to this project is given.

2.2 Causal inference

Inferring the relationship between a cause and its effect is among the most fundamental questions in science. In fact, it traditionally exceeded the scientific domain; historically, the study of causality has been a subject of philosophical debate De Pierris and Friedman (2018).

Specifically, while the philosophical study of causal reasoning dates back to Aristotle Falcon (2019), it was not until the 20$^{\text{th}}$ century when the foundations of causality as a scientific discipline were established.

During the first half of the 20$^{\text{th}}$ century, the work of Sewall Wright in structural equation modelling Wright (1921), of Ronald Fisher in the design of experiments Fisher (1949), or of Bradford Hill in randomized clinical trials Hill (1965), were some of the cornerstones that inspired the development of causal inference, in an effort to advance science from *association* to *causation*. Modern theories of causality emerged in the late 20$^{\text{th}}$ century. Notable examples include the potential outcomes framework Rubin (1974) (and its independent precursor Neyman (1923)), the theory of structural causal models Pearl (2000), and the sufficient cause model Rothman (1976). For a unified causal language as proposed in Pearl (2000), the notion of an *intervention* in a system is of fundamental importance.

For the goals of the project, the focus is on causal inference methods that study *causal relations* between time-dynamic processes, or, alternatively, aim to unveil the *causal structure* of a time-dynamic dataset with interacting variables. As we will see below, in this context, *causality* is

generally assigned a specific meaning, and intervening in a system is not required for inferring causation. These remarks clearly indicate the subset of causality theory to be examined: causal inference in the analysis of time series.

2.2.1 Granger causality

Introducing any method for causal inference implicitly presumes the existence of a concrete definition of causality. For time series analysis, the central notion of causality is the one formalized in Granger (1969) inspired by the ideas of Wiener (1956).

Since the introduction of Granger causality (GC), researchers have introduced other notions of causality in the context of time series, by extending Granger causality or adapting the ideas of other causal inference frameworks to time series Eichler (2012). It is however without a doubt that GC has been the most influential and popular causality concept for time series and a concise overview of it follows.

The intuition behind GC is an improvement in prediction, as envisioned in Wiener (1956):

"For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one."

Granger formalized this concept, postulating the following:

- the cause precedes the effect
- the cause contains information about the effect that is unique, and is in no other variable

According to Granger, a consequence of these two statements is that the causal variable helps in forecasting the effect variable after other data has been first used Granger (2004). While the first statement above is commonly accepted throughout causal inference, the second statement is more subtle as it requires the information provided by X about Y to be unique and separated from all other possible sources of information Eichler (2012). These statements enabled Granger to consider two information "sets", relating to a time series $Y = Y_t$:

- $\mathcal{I}^*(t)$ is the set of "all information in the universe up to time t"
- $\mathcal{I}^*_{-Y}(t)$ contains the same information except for the values of series Y up to time t.

From the discussion above, it is now expected that if Y causes X the conditional distributions of X_{t+1} given the two information sets $\mathcal{I}^*(t)$, $\mathcal{I}^*_{-Y}(t)$ differ from each other.

In other words, Y is said to cause X if Granger (1980):

$$P\Big(X_{t+1} \in A \Big| \mathcal{I}^*(t)\Big) \neq P\Big(X_{t+1} \in A \Big| \mathcal{I}^*_{-Y}(t)\Big) \tag{2.29}$$

Otherwise, if the two probability distributions above are equal, Y does not cause X. Granger causality is then formulated as a statistical hypothesis, with the null hypothesis being equality of distributions and therefore no causation.

While intuitive, (2.29) is more of a concept than a rigorous definition. It is clear that the aforementioned sets $\mathcal{I}^*(t)$, $\mathcal{I}^*_{-Y}(t)$ are not well-defined. Granger himself notes Granger (1980):

"The ultimate objective is to produce an operational definition, which this is certainly not, by adding sufficient limitations."

For mathematical rigor, a specific implementation of this idea is required. Indeed, testing this hypothesis can be done in a variety of ways, from a parametric or non-parametric standpoint, and multivariate extensions have been proposed. Each implementation features its own theory and

results coming from the wider framework it belongs to (see Hlavackova-Schindler et al. (2007) and references therein).

In his initial formulation, Granger implemented this idea within the framework of linear (auto)regression. Consider the following two nested models where $\varepsilon_t, \tilde{\varepsilon}_t$ are the model residuals:

$$X_t = \sum_{i=1}^{q} a_i X_{t-i} + \varepsilon_t \tag{2.30}$$

$$X_t = \sum_{i=1}^{q} a_i X_{t-i} + \sum_{i=1}^{q} b_i Y_{t-i} + \tilde{\varepsilon}_t \tag{2.31}$$

There are now two approaches in this context for inferring Granger causality from source Y to target X, which are roughly equivalent (Bossomaier et al., 2016, Chapter 4):

First, Y is inferred to cause X whenever the full model that includes Y yields a better prediction of X compared to the reduced model that does not. Standard linear prediction theory Hamilton (1994) suggests measuring this by comparing the variances of the residuals $\tilde{\varepsilon}_t, \varepsilon_t$ of the models through their ratio. Following Geweke (1982), the corresponding test statistic is:

$$\mathcal{F}_{Y \to X} = \log \frac{Var(\varepsilon_t)}{Var(\tilde{\varepsilon}_t)} \tag{2.32}$$

The second approach is based on maximum likelihood (see Section 2.4.3). Geweke (1982) notes that, if the residuals $\varepsilon_t, \tilde{\varepsilon}_t$ are normal, $\mathcal{F}_{Y \to X}$ is the log-likelihood ratio test statistic for the model (2.31) under the null hypothesis

$$H_0 : b_1 = b_2 = ... = b_q = 0 \tag{2.33}$$

Recalling (2.29), note that H_0 is equivalent with no Granger causation, since failing to reject H_0 is equivalent with the two information sets $\mathcal{I}^*(t)$ and $\mathcal{I}^*_{-Y}(t)$ being equal.

The estimation of the parameters of the model, including the variance of the residuals, can be achieved through a standard ordinary least squares approach (see Section 2.4.2). Then, the estimator of the test statistic $\hat{\mathcal{F}}_{Y \to X}$ can be calculated.

Since $var(\varepsilon_t) \geq var(\tilde{\varepsilon}_t)$, it holds that $\mathcal{F}_{Y \to X} \geq 0$. Geweke (1982) utilizes large-sample theory to characterize the distribution of the estimator $\hat{\mathcal{F}}_{Y \to X}$ as a χ^2 distribution under the null hypothesis $\mathcal{F}_{Y \to X} = 0$, and a non-central χ^2 distribution under the alternative $\mathcal{F}_{Y \to X} > 0$. Assuming enough data, the appropriate χ^2 distribution is subsequently used to infer about the hypothesis.

An interesting extension to GC was given in Geweke (1984). There, *conditional* Granger causality is introduced. Using the same linear regression framework as before, the time series $Z = Z_t$ is also introduced, which can be thought of as the *side information* in a system. The models (2.30), (2.31) are subsequently expanded by adding the side information Z as an explanatory variable:

$$X_t = \sum_{i=1}^{q} a_i X_{t-i} + \sum_{i=1}^{q} c_i Z_{t-i} + \varepsilon_t \tag{2.34}$$

$$X_t = \sum_{i=1}^{q} a_i X_{t-i} + \sum_{i=1}^{q} c_i Z_{t-i} + \sum_{i=1}^{q} b_i Y_{t-i} + \tilde{\varepsilon}_t \tag{2.35}$$

Then, the existence of conditional Granger causality $Y \to X|Z$ is tested as before:

$$\mathcal{F}_{Y \to X|Z} = \log \frac{var(\varepsilon_t)}{var(\tilde{\varepsilon}_t)} \tag{2.36}$$

2.2.2 Transfer entropy and causality

In this section, the connection between transfer entropy and Granger causality is established and discussed. This is partially achieved through the example of normally distributed variables.

From the discussions before, subtle similarities between transfer entropy and Granger causality already appear. For example, both notions disregard in their definition one of the essential requirements for establishing any causal relation in the traditional sense: that of interventions.

Moreover (see Wiener's original idea in Section 2.2.1), GC is defined in terms of prediction improvement: a Granger-causal relation from Y to X is the degree to which Y *predicts* the future of X beyond the degree to which X already *predicts* its own future.

On the other hand (see discussion below (2.21)), TE is defined in terms of resolution of uncertainty: the transfer entropy from Y to X is the degree to which Y *disambiguates* the future of X beyond the degree to which X already *disambiguates* its own future Barnett et al. (2009).

Barnett et al. (2009) established a rigorous connection between TE and GC by proving the following result, concentrating on the conditional case as formulated in (2.22) and (2.36):

Theorem 2.2.1. *Let $\mathcal{F}_{Y\to X|Z}$ as in (2.36). For three jointly Gaussian and stationary time series*[1] *X_t, Y_t, Z_t it holds that*

$$\mathcal{F}_{Y\to X|Z} = 2T_{Y\to X|Z} \tag{2.37}$$

Furthermore, it was later proved by Serès et al. (2016) that inequality still holds even without the normality assumption:

Theorem 2.2.2. *For three jointly distributed and stationary time series X_t, Y_t, Z_t it holds that*

$$\mathcal{F}_{Y\to X|Z} \leq 2T_{Y\to X|Z} \tag{2.38}$$

The connection between TE and GC is further extended (within the autoregressive framework) to various generalized Gaussian/exponential distributions Schindlerova (2011) and ultimately to a general class of Markov models in a maximum likelihood framework Barnett and Bossomaier (2012). For a more elaborate presentation of the relationship between TE and GC, we refer to (Bossomaier et al., 2016, Section 4.4).

Information Transfer and Causality

At a certain point, results such as those presented in Section 2.2.2 may lead to confusion regarding the differences between transfer entropy and Granger causality. Moreover, the interpretation of transfer entropy as a non-linear and non-parametric *extension* of Granger causality that is popular in the scientific community might exacerbate this problem.

Section 2.2.1 elaborates on what causality actually means, in the context of Granger causality. It is therefore clear that causality in the Granger sense is essentially an improvement in prediction, or a predictive *transfer*. This notion of causality might differ from more traditional causality theories (e.g. Pearl (2000)); but it is intuitive, able to implemented simply through linear models and therefore convenient for practical purposes.

If TE is thought of as an extension of GC (because of results such as those presented in this section), intuitively one might think that the causal content of GC is also extended to TE; making TE a general tool for capturing causality in the predictive transfer sense. This perspective considered by itself can be precarious, as it disregards the theoretical framework that TE ultimately comes from: information theory.

Moreover, besides the predictive transfer sense, causal inference in general is fundamentally associated with *causal effects*. In this sense, causality refers to the source having a *direct influence* in the (next state of) the target, and changes in the target being driven by changes in the source.

As seen in its introductory Section 2.1.3, TE is fundamentally a measure that quantifies the *directed information transfer* from a source to a target.

[1] Defined in Section 2.3

The question now is whether the concept of information transfer is closer to that of predictive transfer (as seen in Granger causality) or causal effect (in the "direct influence" sense). It is thus important to disambiguate the relation of information transfer and causality.

Lizier et al. (2008) state that the relation of these concepts has not been made clear, leading to researchers frequently misusing them by utilizing one to infer about another or even directly equating them. They furthermore argue that the concepts of predictive transfer and causal effect are distinct. Among the two, they assert that the notion of information transfer is closer to that of predictive transfer, and therefore TE is indeed a sensible quantification of causality in the predictive transfer sense. For an information theoretic treatment of causality in the sense of causal effects and direct influences, they propose the measure of *information flow* that was introduced in Ay and Polani (2008) as a more fitting quantification of that notion.

The theoretical presentation of TE concludes with referring to its shortcomings. In an insightful paper, James et al. (2015) demonstrate inherent limitations of TE stemming from the nature of mutual information that have led to misinterpretations. Under specific conditions, TE might overestimate the information flow, or completely miss it. This relates to how information can be decomposed Williams and Beer (2010), and is an active area of research Finn and Lizier (2020).

2.3 Time series

This section includes key notions from the field of time series that are of central importance to the project. Theoretical concepts as well as important examples of time series are presented. This section is largely based on Brockwell and Davis (2009) and Brockwell and Davis (2010).

2.3.1 Stationarity in time series

Stationarity is an important concept that is assumed for many time series analysis methods. However, in real data, stationarity is not always encountered, and non-stationary patterns can contain information that is of utmost importance. This section therefore introduces the concept of stationarity for time series.

Definition 2.3.1 (Autocovariance function)**.** *For a time series* $\{X_t, t \in \mathbb{Z}\}$ *such that* $Var(X_t) < \infty$ *for each* $t \in T$*, the autocovariance function* $\gamma_X(\cdot,\cdot)$ *of* X_t *is defined as:*

$$\gamma_X(t,s) = Cov(X_t, X_s) = E\big[(X_t - E[X_t])(X_s - E[X_s])\big], \quad t,s \in \mathbb{Z} \tag{2.39}$$

Definition 2.3.2 (Weak Stationarity)**.** *The time series* $\{X_t, t \in \mathbb{Z}\}$ *is weakly stationary if*

- $E\big[|X_t|^2\big] < +\infty$ *for all* $t \in \mathbb{Z}$
- $E[X_t] = m$*, for all* $t \in \mathbb{Z}$ *where* $m \in \mathbb{R}$
- $\gamma_X(t,s) = \gamma_X(t+h, s+h)$ *for all* $t,s,h \in \mathbb{Z}$

So, a weakly stationary time series has a finite second moment everywhere, a constant first moment everywhere, and its autocovariance function is invariant under translations. In literature, weak stationarity is also known as covariance stationarity, second order stationarity, or stationarity in the wide sense. For simplicity, throughout the report, the use of the term "stationarity" alone will refer to weak stationarity, and strict stationarity as defined below will be always made explicit.

It is easy to see that the stationarity property implies that $\gamma_X(t,s) = \gamma_X(t-s,0)$. It is therefore convenient to redefine the autocovariance function for stationary time series as a function of one variable (the length of the time interval $t-s$ considered):

$$\gamma_X(h) \equiv \gamma_X(h,0) = Cov(X_{t+h}, X_t) \tag{2.40}$$

In that case, the *autocorrelation* function can also be defined similarly:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} \tag{2.41}$$

Definition 2.3.3 (Strict Stationarity)**.** *The time series $\{X_t, t \in \mathbb{Z}\}$ is strictly stationary if, for any $k \in \mathbb{N}$ and $t_1, ..., t_k, h \in \mathbb{Z}$ the following random vectors have the same distribution:*

$$(X_{t_1}, ..., X_{t_k}) \stackrel{d}{=} (X_{t_1+h}, ..., X_{t_k+h}) \tag{2.42}$$

In other words, if the time series X_t is strictly stationary the distribution of any random vector is invariant under time translations.

It is intuitively expected that strict stationarity implies weak stationarity. This is not exactly right as a time series can be strictly stationary with an infinite second moment, and thus not weakly stationary. But if finiteness is assumed for the second moment of a strictly stationary process, then weak stationary is indeed implied.

Theorem 2.3.4. *A strictly stationary time series $\{X_t, t \in \mathbb{Z}\}$ with $E\left[|X_t|^2\right] < \infty$ for all $t \in \mathbb{Z}$ is weakly stationary.*

Proof. The proof can be found in Appendix A. □

Weak stationarity does not imply strict stationarity in general, and an example of that is also given in Appendix A. There is, however, an important case where that happens: Gaussian time series. Since they are essential for the project, a short introduction to them is given in the next section.

2.3.2 Important examples

In this section, a variety of examples of important time series is introduced. Different concepts of *noise*, the *random walk*, the *autoregressive* process as well as *Gaussian* time series are introduced.

IID noise

A first trivial example of a time series is the i.i.d. noise.

Definition 2.3.5 (i.i.d. noise)**.** *Let X_t be a sequence of independent and identically distributed random variables, with mean zero and variance σ^2. This time series is referred to as i.i.d noise.*

Provided that $E[X^2] = \sigma^2 < \infty$, i.i.d. noise is stationary, with

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases} \tag{2.43}$$

Random Walk

The random walk is obtained by considering the partial sums of i.i.d noise.

Definition 2.3.6 (Random Walk)**.** *A random walk with zero mean is obtained by defining $S_0 = 0$ and letting*

$$S_t = X_1 + X_2 + ... + X_t, \quad \text{for } t = 1, 2, ... \tag{2.44}$$

where X_t are i.i.d random variables.

It holds that $E[S_t] = 0$, $E[S_t^2] = t\sigma^2 < \infty$ for all t and for $h \geq 0$,

$$\gamma_s(t+h, t) = Cov(S_{t+h}, S_t) = Cov(S_t + X_{t+1} + ... + X_{t+h}, S_t) = Cov(S_t, S_t) = t\sigma^2 \tag{2.45}$$

Since $\gamma_s(t+h, t)$ depends on t, the random walk S_t is not stationary.

White Noise

A time series with uncorrelated zero mean random variables is referred to as white noise.

Definition 2.3.7 (White noise)**.** *The time series X_t is called white noise if $E[X_t] = 0$ and $Cov(X_t, X_s) = 0$ for $t \neq s$.*

White noise is clearly stationary, having the same autocovariance function with i.i.d noise. It also holds that every i.i.d noise is white noise, but not conversely.

AR(1)

A very important example of time series is the autoregressive process of order 1, written shortly as AR(1).

Definition 2.3.8. *A first-order autoregressive process X_t is defined recursively as follows:*

$$X_t = \varphi X_{t-1} + Z_t \tag{2.46}$$

where $|\varphi| < 1$, Z_t is a white noise process with variance σ^2 and Z_t is uncorrelated with X_s, for each $s < t$. Here, $t \in \mathbb{I}$ where $\mathbb{I} = \mathbb{N}$ or $\mathbb{I} = \mathbb{Z}$.

For the condition for ϕ we refer to (Brockwell and Davis, 2009, p.81). The index t of an AR(1) process X_t may be defined over $\mathbb{Z}$ or $\mathbb{N}$. In the following, we will contrast these two approaches. The concepts of *stability* and *stationarity* for AR(1) processes warrant separate treatments. For this purpose we introduce the *lag operator*.

Definition 2.3.9. *(Lag operator) Let $\{X_t\}_{t\in\mathbb{Z}}$ be a time series, $k \in \mathbb{Z}$. The lag operator L^k is defined as:*

$$L^k X_t = X_{t-k} \tag{2.47}$$

In the case were $k = 1$, the lag operator maps a value of the time series to the one before it, and the term *backshift operator* B is preferred. Applying the operator $(I - B)$ (where I is the identity operator) to a time series X_t is of particular importance.

Definition 2.3.10. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a time series. The first difference operator Δ is defined as:*

$$\Delta X_t = (I - B)X_t = X_t - X_{t-1} \tag{2.48}$$

The time series $\{\Delta X_t\}_{t\in\mathbb{Z}}$ comprises the (lag-one) increments of X_t. In case $\{\Delta X_t\}_{t\in\mathbb{Z}}$ is stationary we then say that $\{X_t\}_{t\in\mathbb{Z}}$ has stationary increments.

This definition is directly extended for $d \in \mathbb{N}$ to d-order differencing via $\Delta^d X_t := (I - B)^d X_t$.

Then, an AR(1) process is rewritten as:

$$X_t = \varphi X_{t-1} + Z_t \iff \tag{2.49}$$
$$(I - \varphi B)X_t = Z_t \iff \tag{2.50}$$

Obtaining an explicit expression for X_t is now achieved by inverting the operator $(I - \varphi B)$. This happens if and only if $|\varphi| < 1$, in which case $(I - \varphi B)^{-1} = \sum_{i=0}^{\infty} \varphi^i B^i$. This is what we refer to as the *stability* condition for AR(1) processes; this condition for φ was part of the Definition 2.3.8 to ensure that X_t has this representation.

Remark. *An informal explanation of why this holds is given by the geometric series, where the inverse of the number $1 - r$ is equal to $\sum_{i=0}^{\infty} r^i$ if and only if $|r| < 1$. For the analogous result in function spaces that is needed here, we refer to Brockwell and Davis (2009)[Example 3.1.2].*

Thus,

$$X_t = (I - \varphi B)^{-1} Z_t = \sum_{i=0}^{\infty} \varphi^i B^i Z_t = \begin{cases} \sum_{i=0}^{\infty} \varphi^i Z_{t-i} & \text{, if } t \in \mathbb{Z} \\ \sum_{i=0}^{t} \varphi^i Z_{t-i} & \text{, if } t \in \{0, 1, 2, ...\} \end{cases} \tag{2.51}$$

Let $h \in \mathbb{Z}$ and note:

$$X_{t+h} = \sum_{i=0}^{\infty} \varphi^i Z_{t+h-i} = \sum_{i=0}^{h-1} \varphi^i Z_{t+h-i} + \sum_{i=h}^{\infty} \varphi^i Z_{t+h-i} \tag{2.52}$$

$$= \sum_{i=0}^{h-1} \varphi^i Z_{t+h-i} + \sum_{j=0}^{\infty} \varphi^{j+h} Z_{t-j} = \sum_{i=0}^{h-1} \varphi^i Z_{t+h-i} + \varphi^h \sum_{j=0}^{\infty} \varphi^j Z_{t-j} \tag{2.53}$$

$$= \sum_{i=0}^{h-1} \varphi^i Z_{t+h-i} + \varphi^h X_t \tag{2.54}$$

Computing the autocovariance function yields:

$$\gamma_X(h) = Cov(X_{t+h}, X_t) = Cov\big(\sum_{i=0}^{h-1} \varphi^i Z_{t+h-i} + \varphi^h X_t, X_t\big) \tag{2.55}$$

$$= Cov\big(\sum_{i=0}^{h-1} \varphi^i Z_{t+h-i}, X_t\big) + Cov\big(\varphi^h X_t, X_t\big) = \sum_{i=0}^{h-1} Cov\big(\varphi^i Z_{t+h-i}, X_t\big) + Cov\big(\varphi^h X_t, X_t\big) \tag{2.56}$$

$$= 0 + \varphi^h Cov(X_t, X_t) = \varphi^h Var(X_t) \tag{2.57}$$

The variance can be computed from (2.51):

$$Var(X_t) = Var\big(\sum_i \varphi^i Z_{t-i}\big) = \sum_i (\varphi^i)^2 Var\big(Z_{t-i}\big) = \sum_i (\varphi^2)^i \sigma^2 = \sigma^2 \sum_i (\varphi^2)^i \tag{2.58}$$

$$= \begin{cases} \dfrac{\sigma^2}{1 - \varphi^2} & \text{, if } t \in \mathbb{Z} \\ \sigma^2 \dfrac{1 - (\varphi^{t+1})^2}{1 - \varphi^2} & \text{, if } t \in \{0, 1, 2, ...\} \end{cases} \tag{2.59}$$

We then conclude

$$\gamma_X(h) = \begin{cases} \varphi^h \dfrac{\sigma^2}{1 - \varphi^2} & \text{, if } t \in \mathbb{Z} \\ \varphi^h \sigma^2 \dfrac{(1 - (\varphi^{t+1})^2)}{1 - \varphi^2} & \text{, if } t \in \{0, 1, 2, ...\} \end{cases} \tag{2.60}$$

Hence we infer that, for the case where the index is defined over $\mathbb{N}$ (i.e. the process starts from 0), the dependence of the autocovariance function on t implies that the process is always non-stationary. Therefore, in order to allow for the possibility of stationarity when studying an AR(1) process, we will consider time indices over all integers, i.e. "starting" from $-\infty$. This can still be combined with interpreting the t parameter as time, by assuming that the index was defined over all integers but we only observed its values from $t = 0$. In that case, X_0 is a random variable (Kirchgässner and Wolters, 2007, Section 2.1). The AR(1) system is then stationary if and only if $|\varphi| < 1$. Notice that for these AR(1) models, stability and stationarity conditions coincide.

Gaussian Time Series

In defining Gaussian time series, the multivariate normal distribution should be introduced and discussed first. A comprehensive presentation of multivariate normality is therefore included in Appendix A. We thus proceed with defining Gaussian time series:

Definition 2.3.11. *A time series $\{X_t\}_{t\in\mathbb{Z}}$ is a Gaussian time series if any finite vector of random variables $(X_{t_1}, ..., X_{t_k})$ is multivariate normally distributed.*

For the goals of the project, a great advantage offered by Gaussian time series is their flexibility, since due to their multivariate normal structure, Gaussian time series are fully specified by their mean $\mu(t) = E[X_t]$ and autocovariance function $\kappa(s,t) = Cov(X_s, X_t)$. Their stationarity properties also depend on them.

Moreover, for Gaussian time series weak stationarity implies strict stationarity. Indeed, if a Gaussian time series is weakly stationary, then for all $n = 1, 2, ...$ and for all $h, t_1, t_2, ..., t_n \in \mathbb{Z}$ the random vectors $(X_{t_1}, ..., X_{t_n})$ and $(X_{t_1+h}, ..., X_{t_n+h})$ have the same mean and covariance matrix, and hence the same (multivariate normal) distribution. In fact, since Gaussian processes have a finite variance everywhere, strict and weak stationarity are fully equivalent for them.

Gaussian time series are important in time series analysis, as they allow for convenient probabilistic calculations. This is also the case for entropy and other information theoretic quantities, to be studied later.

2.4 Estimation techniques

In this chapter, the focus so far has mostly been on probabilistic aspects of the theories involved. In this section the focus shifts towards estimation: several well-known estimation techniques that are relevant to the project are shortly presented.

2.4.1 Density estimation

This section is devoted to the estimation of probability densities. Before introducing the theory, it is important to disambiguate the notation and terminology that is used - for the discrete and continuous case. Recall that Shannon entropy $H(X)$ is associated with discrete random variables and probability mass functions, while differential entropy $h(X)$ is associated with continuous random variables and probability densities.

Entropy estimation is of central importance to this project, and from the above discussion we note that Shannon entropy estimation entails estimating functions of probability mass functions while differential entropy estimation requires the estimation of probability densities. Generally, the discrete case is more convenient as we can directly estimate $p_X(x)$ by simply calculating the corresponding relevant frequencies in our data. In the continuous case, a standard approach to density estimation is kernel density estimators Silverman (1986). An overview is given below, starting from an intuitive and well-known method inspired by the discrete case: the histogram.

Kernel density estimation

A widely used density estimator is the histogram. Let $X_1, \ldots, X_n$ be a random sample from a distribution function F with continuous derivative $F' = f$.

Definition 2.4.1 (Empirical distribution function)**.** *For a set $A \subseteq \mathbb{R}$, the empirical distribution function P_n is defined as follows:*

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A \tag{2.61}$$

So, the empirical distribution function gives the number of elements in the sample that belong to the set A, divided by the (current) sample size. When, for an arbitrary $t \in \mathbb{R}$, $A = (-\infty, t)$ the

empirical distribution function is a natural estimator for the distribution function F, and possesses various favorable properties (van der Vaart, 1998, Section 19.1).

Let I be a compact interval on $\mathbb{R}$ and suppose that the intervals $I_1, \ldots, I_k$ form a partition of I, i.e.

$$I = I_1 \cup \ldots \cup I_k, \quad I_i \cap I_j = \emptyset \text{ if } i \neq j.$$

Definition 2.4.2 (Histogram)**.** *The histogram of $X_1, \ldots, X_n$ with respect to the partition $I_1, \ldots, I_k$ is defined as*

$$H_n(x) = \sum_{j=1}^{k} \frac{P_n(I_j)\, \mathbb{1}_{I_j}(x)}{|I_j|} \tag{2.62}$$

where $|I_j|$ denotes the length of the interval I_j.

It is clear that the histogram is a stepwise constant function. Two major disadvantages of the histogram that impact its capabilities as a density estimator are:

- the stepwise constant nature of the histogram
- the fact that the histogram heavily depends on the choice of the partition

It is because of this phenomenon that histograms are not recommended as a density estimator. A natural way to improve on histograms is to get rid of the fixed partition by putting an interval around each point. If $h > 0$ is fixed, then

$$\widehat{N}_n(x) = \frac{P_n((x-h, x+h))}{2h} \tag{2.63}$$

is called the *naive density estimator* and was introduced in 1951 by Fix and Hodges (reprinted in Fix and Hodges (1989)). The motivation for the naive estimator is that

$$P(x-h < X < x+h) = \int_{x-h}^{x+h} f(t)\, dt \approx 2\, h\, f(x). \tag{2.64}$$

It is intuitively clear from (2.64) that the bias of $\widehat{N}_n$ decreases as h tends to 0. However, if h tends to 0, then one is using less and less observations, and hence the variance of $\widehat{N}_n$ increases. The optimal value of h is a compromise between the bias and the variance.

The naive estimator is a special case of the following class of density estimators. Let K be a *kernel function*, that is a non-negative function such that

$$\int_{-\infty}^{\infty} K(x)\, dx = 1. \tag{2.65}$$

Definition 2.4.3 (Kernel estimator)**.** *The kernel estimator with kernel K and bandwidth h is defined by*

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \tag{2.66}$$

Thus, the kernel indicates the weight that each observation receives in estimating the unknown density. All kernel estimators are densities, and the naive estimator is a kernel estimator with kernel

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Examples of popular kernels in research and applications are given in Table 2.1. Avoiding to fix the bandwidth h, a nearest neighbor density estimator may be utilized. As we will later see, nearest neighbor estimation is particularly important for Information Theory.

Definition 2.4.4 (kNN density estimation)**.** *The k-nearest neighbor kernel estimator with kernel K is defined by*

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_k(x)}K\left(\frac{x - X_i}{d_k(x)}\right) \tag{2.67}$$

where $d_k(x)$ is the distance of x from its k^{th} nearest neighbor in the dataset.

Kernel name	Gaussian	Naive/Rectangular	Triangular	Epanechnikov
Function	$\frac{1}{\sqrt{2\,\pi}}\,e^{-\frac{1}{2}\,x^2}$	$\frac{1}{2}\,I_{(-1,1)}(x)$	$(1-\lvert x\rvert)\,I_{(-1,1)}(x)$	$\frac{3}{4}\,(1-x^2)\,I_{(-1,1)}(x)$

Table 2.1: Popular kernels for density estimation

2.4.2 Least squares estimation

Least squares estimation is a general parametric method for deriving point estimators of population parameters. One of its main applications, is the estimation of the parameters of a model that best fit a dataset. The fit of a model to a dataset is measured by its residuals, i.e. the difference between the actual value of the response variable and the value that the model predicts. By minimizing the sum of squared residuals of the model this procedure calculates the optimal model parameters.

Least squares methods fall into two categories: linear and non-linear. Since the inclusion of least squares theory in this section is motivated by the importance of Granger causality (which was traditionally implemented within linear regression) we concentrate on linear least squares.

Similarly, linear least squares methods consist of several variants, such as ordinary, weighted or generalized linear least squares methods. Here we focus on the simplest of these methods: ordinary (linear) least squares.

Ordinary Least squares

Consider the datapoints $(x_{i1}, y_i), ..., (x_{im}, y_i)$, $i = 1, ..., n$, and $m \leq n$, where X_i are the independent variables and Y is the dependent variable. We want to find the best fitting straight line which, given a set of values of the independent variables X_i, returns a value of the dependent variable Y.

$$y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_m x_{im} \tag{2.68}$$

Definition 2.4.5 (OLS)**.** *The ordinary least squares estimates $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m$ are defined to be the values of $\theta_1, \theta_2, ..., \theta_m$ that minimize the sum of squared residuals of the model*

$$S(\theta_1, ..., \theta_m) = \sum_{i=1}^{n}(y_i - \theta_1 x_{i1} - ... - \theta_m x_{im})^2 \tag{2.69}$$

Treating this as a minimization problem, and setting the partial derivatives of S with respect to $\theta_1, ..., \theta_m$ equal to zero shows that the vector $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)^T$ that minimizes (2.69) is the solution of the *normal equations*

$$X^T X\hat{\theta} = X^T y \tag{2.70}$$

where X is the n x m matrix $X = [x^{(1)}, ..., x^{(m)}]$ and $x^{(i)}$ is the column vector $x^{(i)} = (x_{i1}, ..., x_{im})^T$. The solution of this equation is unique iff $X^T X$ is non-singular (a condition that is equivalent with X having full rank), in which case

$$\hat{\theta} \equiv (\hat{\theta}_1, ..., \hat{\theta}_m) = (X^T X)^{-1} X^T y \tag{2.71}$$

2.4.3 Maximum likelihood estimation

Maximum likelihood is another general parametric approach for the estimation of population parameters.

Definition 2.4.6 (Maximum Likelihood function)**.** *Consider a random variable X and a parametric form for its density $f(x;\theta_1,...,\theta_r)$ where $\theta_1,...,\theta_r$ are unknown population parameters. Given an i.i.d sample $X_1,..,X_n$ the likelihood function is defined as:*

$$L(\theta_1,...,\theta_r;X_1,...,X_n)=\prod_{i=1}^{n}f(X_i;\theta_1,...,\theta_r) \tag{2.72}$$

The likelihood function $L(\theta_1,...,\theta_r;X_1,...,X_n)$ is thought of as a function of the unknown parameters $\theta_1,...,\theta_r$ only. The *maximum likelihood estimators* (MLE's) $\hat{\theta}_1,...,\hat{\theta}_r$ of $\theta_1,...,\theta_r$ are the values that maximize the likelihood function $L(\theta_1,...,\theta_r;X_1,...,X_n)$.

This optimization problem is frequently transformed by working with the logarithm of (2.72) (when it is defined) and therefore transforming the product to a sum. In that case the resulting function is known as the *log-likelihood* function, and the MLE's $\hat{\theta}_1,...,\hat{\theta}_r$ of $\theta_1,...,\theta_r$ are the solution of the following maximization problem:

$$\max_{\tilde{\theta}_1,...,\tilde{\theta}_r}\left[\log L(\theta_1,...,\theta_r;X_1,...,X_n)\right]=\max_{\tilde{\theta}_1,...,\tilde{\theta}_r}\sum_{i=1}^{n}\log\left[f(X_i;\tilde{\theta}_1,...,\tilde{\theta}_r)\right] \tag{2.73}$$

Due to the logarithm being a monotonic function, the value that maximizes (2.73) is the same with the value that maximizes (2.72). This can again be treated as a maximization problem, although a closed-form solution for the problem will rarely be attainable - and numerical techniques are employed. Maximum likelihood estimators possess a variety of favorable asymptotic properties.

Likelihood-ratio test

The MLE framework is closely related to hypothesis testing, since a null hypothesis is frequently expressed by theorizing that the parameters $\theta=(\theta_1,...,\theta_r)$ of a model belong to a subset Θ_0 of the parameter space Θ, while the alternative hypothesis assumes them to belong to its complement $\Theta\setminus\Theta_0$. This hypothesis can subsequently be tested through the *log-likelihood ratio test statistic* that is given by

$$\lambda=-2\log\left[\frac{\sup_{\theta\in\Theta_0}L(\theta)}{\sup_{\theta\in\Theta}L(\theta)}\right] \tag{2.74}$$

Chapter 3

Estimating Entropy

Following the establishment of various probabilistic aspects of information theory before, we now turn to estimation. Even if entropy was introduced by Shannon in 1948, the estimation of entropy and other information theoretic quantities is still a vibrant field of research. There are several survey papers documenting different approaches and advances for estimation in information theory such as Beirlant et al. (1997), Paninski (2003) and Hlavackova-Schindler et al. (2007). A modern presentation that is relevant to transfer entropy specifics is featured in Bossomaier et al. (2016), which is what we generally follow in this chapter.

3.1 Entropy estimators

It is clear that if an i.i.d sample $X_1, ..., X_n$ is assumed, then any technique that estimates the probability function p_X of X is also an entropy estimator - by simply substituting the estimator in the definition of entropy. These estimators of entropy are the first to be discussed.

3.1.1 Plug-in estimators

Considering the discrete case, recall the definition of Shannon entropy of a random variable X:

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \tag{3.1}$$

An estimator for Shannon entropy can indeed be immediately provided through the estimation of the corresponding probability mass function. This can be conveniently accomplished as described in the beginning of Section 2.4.1. Specifically, for a given set of M bins (that are fixed and chosen independently from the sample) we consider the relative amount of values from the sample that fell in each of them.

Definition 3.1.1 (Shannon entropy plug-in estimator)**.** *Let $x_1, \ldots, x_N$ an i.i.d sample of a discrete random variable X. The plug-in estimator of its Shannon entropy $H(X)$ is defined as*

$$\widehat{H}(X) = -\sum_{i=1}^{M} \widehat{p}_j \log \widehat{p}_j \tag{3.2}$$

where M is the number of bins, $\widehat{p}_j = \frac{n_j}{N}$ for n_j samples in bin j from N samples in total, and $0 \log 0 = 0$.

We hence estimate Shannon entropy by substituting or *plugging-in* each estimate $\widehat{p}_j$ in the place of p_j. This plug-in estimator was extensively studied in Antos and Kontoyiannis (2001). With regards to its bias, it holds that the plug-in estimator underestimates the entropy:

Theorem 3.1.2. *The bias of the plug-in Shannon entropy estimator is bounded by:*

$$-\log\left(1+\frac{M-1}{N}\right) \leq B(\widehat{H}(X)) := E[\widehat{H}(X)] - H(X) \leq 0 \tag{3.3}$$

The lower bound is tight for $N/M \to 0$ and the upper bound is tight for $N/M \to +\infty$.

The proof of Theorem 3.1.2 follows from the non-negativity of a K-L divergence obtained through a variant of the delta method Doob (1935), combined with a result on the upper bound of K-L divergences Gibbs and Su (2002) and Jensen's inequality Paninski (2003). Aiming to compensate this underestimation, Miller (1955) proposed a simple bias correction for the plug-in estimator:

Definition 3.1.3 (Bias correction, Miller)**.** *The compensated Shannon entropy plug-in estimator of Miller is defined as:*

$$\widehat{H}_c(X) = \widehat{H}(X) + \frac{M-1}{N} \tag{3.4}$$

where M and N are defined as in (3.1.1)

Antos and Kontoyiannis (2001) also derived a bound for the variance of the plug-in estimator as well as a concentration inequality both using McDiarmid's inequality McDiarmid (1989):

Theorem 3.1.4 (Variance upper bound)**.** *As defined in* (3.1.1)*, an upper bound for the variance of $\widehat{H}(X)$ is*

$$Var(\widehat{H}(X)) \leq \frac{(\log N)^2}{N} =: v_{max} \tag{3.5}$$

Theorem 3.1.5. *Let $0 < \varepsilon < 1$, v_{max} defined as in* (3.1.4)*. The following inequality holds for the Shannon plug-in estimator $\widehat{H}(X)$:*

$$P(|\widehat{H}(X) - E[\widehat{H}(X)]| > \varepsilon) \leq 2\exp\left\{-\frac{\varepsilon^2}{2v_{max}}\right\} \tag{3.6}$$

Here, the probability that the error of the plug-in estimate is greater than a threshold ε decays exponentially as ε increases. In other words, it is less likely that the plug-in estimate $\widehat{H}(X)$ is going to be far away from its mean $E[\widehat{H}(X)]$, i.e. the values that we generally observe for $\widehat{H}(X)$ are highly concentrated around a mean.

The above results are used in Paninski (2003) to arrive at the bias-variance trade-off for the plug-in entropy estimator. When $N/M \to +\infty$ their ratio takes the following form:

$$\frac{Var(\widehat{H}(X))}{(B(\widehat{H}(X)))^2} = \frac{N(\log M)^2}{M^2} \tag{3.7}$$

We thus observe a linear dependence of the variance/bias ratio on the sample size N, and (since $\log^2 M$ grows much slower than M^2) an inverse quadratic dependence on M. Hence, bias will be generally dominating the estimates when the number of bins M is large, unless the sample size N is huge.

Accounting for this trade-off as well as correcting for small samples Bonachela et al. (2008), propose the following balanced estimator for Shannon entropy:

Definition 3.1.6 (Shannon entropy balanced estimator)**.** *A bias-variance balanced estimator for Shannon entropy is given by:*

$$\widehat{H}_{bal}(X) = \frac{1}{N+2}\sum_{i=1}^{M}\left[(n_i+1)\sum_{j=n_i+2}^{N+2}\frac{1}{j}\right] \tag{3.8}$$

Where, as before, M is the number of bins, N is the sample size, and n_i is the number of occurrences of X in bin i.

Remark. *The straightforward idea of counting relevant frequencies in data (alongside potential bias corrections) to estimate the required probabilities can be applied not only to Shannon entropy, but as long as data are discrete, to mutual information and transfer entropy. This is a simple and fast technique. From now on, the interest is on estimating information theory quantities in the more complicated continuous case.*

Following similar ideas, we can proceed with the continuous case and estimate the differential entropy h of a continuous random variable X with density f.

By estimating the density f we likewise obtain a plug-in estimate for differential entropy. This can be achieved via kernel density estimation as described in Section 2.4.1. The first such estimator was given, for a real random variable X, in Dmitriev and Tarasenko (1974). Given a continuous sample $X_1, \ldots, X_n$, it was proposed to estimate the differential entropy h of X by:

$$\widehat{h}_n(X) = -\int_{A_n} \widehat{f}_n(x) \log \widehat{f}_n(x) dx \tag{3.9}$$

Here, A_n is a symmetric real interval and $\widehat{f}_n(x)$ is a kernel density estimator. Properties of this estimator were studied in Mokkadem (1989) and it was extended to the multivariate case in Joe (1989). Moreover, Grassberger (1988) notes that kernel density based plug-in estimators for differential entropy still contain bias, and tries to correct for it.

The difficulty of evaluating integrals in higher dimensions alongside bias issues indicate that, while convenient, plug-in estimators for differential entropy are of limited use. In the following section, we thus turn to alternative differential entropy estimation methods.

3.1.2 Other estimators

In this section, we discuss two popular estimators of differential entropy. Other important estimators for mutual information and transfer entropy that are presented later are based on them.

Digamma estimator

We start by defining the digamma estimator of differential entropy in the case of real random variables. To do so, the digamma function ψ has to be defined first. Throughout the rest of the chapter, the digamma function will be encountered often, so besides a definition we also provide a few complementary remarks.

Definition 3.1.7 (Digamma function)**.** *The digamma function $\psi : \mathbb{R} \to \mathbb{R}$ is defined as the derivative of the logarithm of the gamma function Γ:*

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{3.10}$$

Utilizing the following well known recursive formula for the gamma function:

$$\Gamma(x+1) = x\Gamma(x) \tag{3.11}$$

we can deduce a similar recursive formula for the digamma function. Differentiating (3.11) with respect to x, dividing by $\Gamma(x+1)$ and re-using (3.11) we get:

$$\frac{\Gamma'(x+1)}{\Gamma(x+1)} = \frac{\Gamma'(x)}{\Gamma(x)} + \frac{1}{x} \iff \tag{3.12}$$

$$\frac{d}{dx} \log(\Gamma(x+1)) = \frac{d}{dx} \log(\Gamma(x)) + \frac{1}{x} \iff \tag{3.13}$$

$$\psi(x+1) = \psi(x) + \frac{1}{x} \tag{3.14}$$

When x is considered over the set of natural numbers, the above result establishes a connection of the digamma values to the harmonic numbers H_n. The following corollary thus also aids our intuition about how the digamma values progress:

Corollary 3.1.8. *For $n \in \{1, 2, 3, \ldots\}$ the digamma function satisfies*

$$\psi(n) = H_{n-1} - C \tag{3.15}$$

where $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is the n^{th} harmonic number, $H_0 := 0$, and C is the Euler-Mascheroni constant: $C = \lim_{n\to+\infty} (H_n - \log n) = 0.57721...$

Returning to differential entropy estimation, we are now able to define the digamma estimator.

Definition 3.1.9 (Digamma estimator)**.** *Let $x_1, x_2, \ldots, x_N$ be an i.i.d real sample of a random variable X. Then, the digamma estimator for the differential entropy $h(X)$ is*

$$\widehat{h}(X) = -\psi(1) + \psi(N) + \frac{1}{N-1} \sum_{i=1}^{N-1} \log \left(x_{(i+1)} - x_{(i)} \right) \tag{3.16}$$

where $x_{(j)}$ denotes the the j^{th} order statistic, i.e. the j^{th} smallest value of the sample.

Kozachenko-Leonenko (K-L) estimator

An important multivariate extension to the digamma estimator was given in Kozachenko and Leonenko (1987). This was achieved by replacing the distances $x(i+1) - x(i)$ between the increasing consecutive points $x(i), x(i+1)$ with nearest-neighbor distances in metric spaces. Here, we restrict the presentation to Euclidean spaces. The Kozachenko-Leonenko estimator of differential entropy is defined as:

Definition 3.1.10 (K-L estimator of differential entropy)**.** *Let $x_1, ..., x_N$ be an i.i.d sample of the random variable X, with $\mathcal{X} \subseteq (\mathbb{R}^d, ||\cdot||)$. Then, the Kozachenko-Leonenko (K-L) estimator of the differential entropy $h(X)$ is given by*

$$\widehat{h}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \varepsilon(i) \tag{3.17}$$

where c_d is the volume of the unit ball of the space $(\mathbb{R}^d, ||\cdot||)$, k is an arbitrary natural number and $\varepsilon(i)$ is twice the distance from x_i to its k^{th} nearest neighbor in the dataset.

Note that the choice of the norm $||\cdot||$ will impact the estimation Table 3.1 summarizes the differences between selecting the maximum vs the Euclidean norm (see Gao et al. (2015) for a discussion of norm selection impact in the KSG estimator). In practice $k = 4$ is recommended as a robust choice for kNN estimation, although this also depends on the sample size.

Norm	Definition	Volume of $B(0,1)$
Maximum	$\|\|x\|\|_\infty := \max_{i=1,\ldots,d}\{x_i\}$	$c_d = 1$
Euclidean	$\|\|x\|\|_2 := \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}$	$c_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$

Table 3.1: The unit ball volume in $\mathbb{R}^d$ for two different norms.

3.2 Mutual information estimators

Having already presented estimators for entropy, we now consider the estimation of mutual information. In Section 3.1 we saw two different categories of entropy estimators. Each category approaches the problem of estimating entropy differently: plug-in estimators directly estimate the probability functions involved, while the digamma and Kozachenko-Leonenko estimators are based on particular geometric features of the samples.

Plug-in estimators

Recalling that $I(X;Y) = H(X) + H(Y) - H(X,Y)$ we observe that marginal and joint entropy estimators, immediately yield a mutual information estimator, while using the analytic expression for MI via the K-L divergence we infer that (multivariate) density estimators also yield a mutual information estimator. Thus, the ideas depicted in Section 3.1 can be extrapolated to mutual information estimation.

In particular, Moon et al. (1995) studies kernel density estimators for mutual information by simply using kernel density estimators for the three probability density functions involved. Such estimators are biased, so the need to correct for it remains (e.g. as in Grassberger (1988) for entropy). However, simultaneous bias correction for individual kernel estimates might prove to have a more adverse effect in the estimate than the bias itself, as noted in Kaiser and Schreiber (2002), and therefore a higher level of complexity for such naive mutual information estimation is introduced.

This is the reason digamma estimators are preferred - and the importance of the multivariate extension provided by the Kozachenko-Leonenko (K-L) estimator is further illustrated by the multivariate nature of mutual information. Starting from the K-L estimator, Kraskov et al. (2003) introduced the *KSG estimator* (named after the authors Kraskov, Stögbauer and Grassberger) that today constitutes the state-of-the-art for MI estimation.

The KSG estimator

Considering the mutual information representation $I(X;Y) = H(X) + H(Y) - H(X,Y)$ we note that the K-L estimator may be immediately used for estimating both marginal entropies $H(X), H(Y)$. For the joint entropy term $H(X,Y)$ we consider the joint random variable $Z = (X,Y)$ that exists in the joint space $\mathcal{X} \times \mathcal{Y}$ equipped with the maximum norm $||\cdot||_\infty$. Therefore, in Definition 3.1.10, d is replaced by $d_Z = d_X + d_Y$, c_d is replaced by $c_{d_X} c_{d_Y}$ and x_i is replaced by $z_i = (x_i, y_i)$. This leads to the following K-L estimate for $h(X,Y)$, where $\varepsilon(i)$ is twice the distance from $z_i = (x_i, y_i)$ to its k^{th} neighbor:

$$\widehat{h}(X,Y) = \psi(k) - \psi(N) - \log c_{d_X} c_{d_Y} - \frac{d_X + d_Y}{N} \sum_{i=1}^{N} \log \varepsilon(i) \tag{3.18}$$

An MI estimator is then directly obtained by substituting all three K-L estimates:

$$\widehat{I}(X;Y) = \widehat{h}(X) + \widehat{h}(Y) - \widehat{h}(X,Y) \tag{3.19}$$

A subtle yet important detail in the above procedure is whether the same number k of nearest neighbors shall be used for all three estimations. Kraskov et al. (2003) noted that doing so will hinder the performance of the estimator and introduced two slightly different MI estimators circumventing this issue. Both estimators rely on not using the same number k of neighbors on the joint and marginal spaces.

Definition 3.2.1 (First KSG MI estimator)**.** *Let X, Y two continuous random variables. The first KSG estimator of the mutual information $I(X;Y)$ is*

$$I^{(1)} = \psi(k) + \psi(N) - \frac{1}{N} \sum_{i=1}^{N} \psi(n_x(i) + 1) - \frac{1}{N} \sum_{i=1}^{N} \psi(n_y(i) + 1) \tag{3.20}$$

where, k is an arbitrary natural number, $\varepsilon_k(i)$ is the distance of z_i to its k^{th} nearest neighbor in the joint $\mathcal{X} \times \mathcal{Y}$ space, N is the amount of samples $z_i = (x_i, y_i)$ in the joint space and $n_x(i), n_y(i)$ are the number of points strictly within a distance $\varepsilon_k(i)$ in the marginal spaces $\mathcal{X}, \mathcal{Y}$ respectively.

The distance threshold $\varepsilon_k(i)$ for the counting that was performed in $n_x(i)$ and $n_y(i)$ in the first KSG estimator was based on the k^{th} nearest neighbor distance on the joint space. Changing this from the joint space to each marginal space yields the second KSG estimator. So, we replace $n_x(i), n_y(i)$ by the number of points in the respective marginal spaces that are located at most $\varepsilon_x(i), \varepsilon_y(i)$ away from x_i, y_i respectively where distances $\varepsilon_x(i), \varepsilon_y(i)$ of $z_i = (x_i, y_i)$ from its k^{th} nearest neighbor are based on the norm of the marginal spaces. The second KSG estimator is:

$$I^{(2)} = \psi(k) + \psi(N) - \frac{1}{k} - \frac{1}{N}\sum_{i=1}^{N}\psi(n_x(i)) - \frac{1}{N}\sum_{i=1}^{N}\psi(n_y(i)) \tag{3.21}$$

3.3 Transfer entropy estimators

Recalling (2.21), transfer entropy is defined as conditional mutual information. It can be subsequently expanded to a linear combination of (joint) entropy terms:

$$T^{(k,\ell)}_{Y\to X}(t) = h(X_t, X^{(k)}_{t-1}) - h(X^{(k)}_{t-1}) - h(X_t, X^{(k)}_{t-1}, Y^{(\ell)}_{t-1}) + h(X^{(k)}_{t-1}, Y^{(\ell)}_{t-1}) \tag{3.22}$$

Similarly to how the KSG method utilized appropriate K-L estimators for each entropy term arising in the definition of mutual information, an analogous KSG method can be used to estimate each term of (3.22), and therefore estimate transfer entropy. The important condition to consider when doing so is choosing appropriate numbers for each nearest neighbor estimation, in the spirit of the KSG estimators described before.

Generally, the discussion at the beginning of Section 3.2 regarding the use of entropy estimators for estimating mutual information applies here too. This can be seen if we also consider the equivalent analytic form of TE:

$$T^{(k,\ell)}_{Y\to X}(t) = \sum_{x_t, x^{(k)}_{t-1}, y^{(l)}_{t-1}} p(x_t, x^{(k)}_{t-1}, y^{(\ell)}_{t-1}) \log \frac{p(x_t | x^{(k)}_{t-1}, y^{(\ell)}_{t-1})}{p(x_t | x^{(k)}_{t-1})} \tag{3.23}$$

Clearly, plug-in estimators such as the kernel density estimators discussed before can be employed in estimating TE via its analytic form. However, following the equivalent discussion for MI estimation, we immediately infer that such approaches will have major drawbacks: high dimensionality was among the main reasons of rejecting kernel density estimators for MI. TE is much more complicated than MI in that regard, reaching arbitrarily high dimensions depending on the choice of the embedding parameters. We therefore restrict our attention to alternative approaches.

KSG estimator

Imitating the KSG method, we apply a K-L estimator based on m^{th} nearest neighbor of each point in the joint space $\mathcal{X}_t \times \mathcal{X}^{(k)}_{t-1} \times \mathcal{Y}^{(\ell)}_{t-1}$ that is again equipped with the maximum norm. If $\varepsilon(i)$ is twice the max-norm distance of the i^{th} point to its m^{th} nearest neighbor in the joint space, we then denote by $n_1(i)$ the neighbor count of the i^{th} point strictly within norm $\varepsilon(i)$ in the $\mathcal{X}^{(k)}_{t-1}$ marginal space, and by $n_2(i)$ and $n_3(i)$ the respective neighbor counts of the i^{th} point strictly within $\varepsilon(i)$ in the smaller joint spaces $\mathcal{X}_t \times \mathcal{X}^{(k)}_{t-1}$ and $\mathcal{X}^{(k)}_{t-1} \times \mathcal{Y}^{(\ell)}_{t-1}$ (that are also equipped with max-norms).

The first KSG estimator defined above subsequently leads to the following estimator for TE:

Definition 3.3.1 (KSG estimator for TE)**.** *For two time series $X_t, Y_t, t \in \mathbb{Z}$, the first KSG*

estimator for the transfer entropy $\widehat{T}^{(k,\ell)}_{Y\to X}(t)$ *is given by*

$$\widehat{T}^{(k,\ell)}_{Y\to X}(t) = I^{(1)}(X_t; Y^{(\ell)}_{t-1}|X^{(k)}_{t-1}) \tag{3.24}$$

$$= \psi(m) - \frac{1}{N}\sum_{i=1}^{N}\big[\psi(n_2(i)) - \psi(n_3(i)) + \psi(n_1(i))\big] \tag{3.25}$$

where m *is an arbitrary natural number, and* $n_1(i), n_2(i), n_3(i)$ *are defined above.*

Similarly, the second KSG estimator for MI yields the second KSG estimator for TE:

$$\widehat{T}^{(k,\ell)}_{Y\to X}(t) = I^{(2)}(X_t; Y^{(\ell)}_{t-1}|X^{(k)}_{t-1}) \tag{3.26}$$

$$= \psi(m) - \frac{2}{m} + \frac{1}{N}\sum_{i=1}^{N}\big[\psi(n_1(i)) - \psi(n_2(i)) + \frac{1}{n_2(i)} - \psi(n_3(i)) + \frac{1}{n_3(i)}\big] \tag{3.27}$$

Since conditional transfer entropy $T_{Y\to X|Z}$ simply adds a second component to the conditional part of the corresponding conditional mutual information $I(X_t; Y^{(\ell)}_{t-1}|X^{(k)}_{t-1}, Z^{(m)}_{t-1})$, the above ideas are trivially extended for estimating conditional TE.

3.4 The non-stationary case

All the estimators of the information theoretic quantities we presented so far rely on strong probabilistic assumptions for the data: for instance, the K-L estimator (as well as the KSG estimator that is based in it) assume i.i.d data, a condition that implies (strict) stationarity in the context of time series. In reality, this assumption is frequently invalidated. Dealing with non-stationarity within Information Theory is an important open question Vu et al. (2008). In this section, we thus review several approaches aimed at dealing with non-stationary data mostly from the perspective of TE estimation.

3.4.1 Data transformations

A very convenient method for resolving non-stationarity throughout the analysis of time series is data transformations. This is a popular workaround that aims to manipulate a non-stationary dataset in such a way that it becomes stationary and to subsequently apply methods that assume stationarity to them (such as the KSG estimator in our context). It is, however, not a panacea; as (Kantz and Schreiber, 2006, Chapter 13) note, only under very specific conditions such data transformations will maintain a theoretical connection to the original dataset.

Differencing & Log transform

The lag operator and first order differencing defined in (2.3.10) already supply a useful transformation method. Considering the time series of the consecutive differences of data may already remove non-stationarities, especially if they are concentrated on the mean of the time series, e.g. in the form of a drift, while higher order differencing will magnify this effect.

Moreover, for time series with positive values, a logarithmic transformation is regarded as helpful in alleviating non-stationarities on the variance of data.

3.4.2 Other methods

This section features alternative methods that aim to deal with the problem of estimation in information theory in the case of non-stationary data. This is achieved by a non-trivial manipulation of data or by making extra assumptions regarding the context.

Symbolic Transfer Entropy

Symbolic transfer entropy (STE) was introduced in Staniek and Lehnertz (2008) for the bivariate case. While this is a theoretical quantity that is distinct from transfer entropy, its main advantage compared to regular TE is related to estimation, so it is shortly introduced in this chapter. It is based on the idea of permutation entropy introduced in Bandt and Pompe (2002). It rank-transforms the data, and technically it can be categorized as another data transformation method. Since STE and its multivariate version have been popular in TE literature (e.g. Ku et al. (2011), Kowalski et al. (2010)) it is separately presented from other common data transformation methods shown before.

Let $X_t, Y_t, t \in \mathbb{Z}$ be two univariate time series, and consider their embedding vectors at an arbitrary time t:

$$X_t^{(\tau_1,d_1)} = (X_t, X_{t-\tau_1}, ..., X_{t-(d_1-1)\tau_1}) \tag{3.28}$$

$$Y_t^{(\tau_2,d_2)} = (Y_t, Y_{t-\tau_2}, ..., X_{t-(d_2-1)\tau_2}) \tag{3.29}$$

For this t, arrange in ascending order all d_1, d_2 values of the embedding vectors of X, Y respectively, using the same ordering as the one already present in the vectors in case of ties. Thus, consider:

$$X_{t-(r_{t,1}-1)\tau_1} \leq X_{t-(r_{t,2}-1)\tau_1} \leq ... \leq X_{t-(r_{t,d_1}-1)\tau_1} \tag{3.30}$$

$$Y_{t-(q_{t,1}-1)\tau_2} \leq Y_{t-(q_{t,2}-1)\tau_2} \leq ... \leq Y_{t-(q_{t,d_2}-1)\tau_2} \tag{3.31}$$

where the $\{r_{t,j},\ j = 1, ..., d_1\}$ are all different, $r_{t,j} \in \{1, ..., d_1\}$ and the equivalent properties for $q_{t,j}$ apply.

Subsequently, define two *symbols* as:

$$\widehat{X}_t := (r_{t,1}, r_{t,2}, ..., r_{t,d_1}) \tag{3.32}$$

$$\widehat{Y}_t := (q_{t,1}, q_{t,2}, ..., q_{t,d_2}) \tag{3.33}$$

For example, the three-dimensional embeddings $(3, 4, 5)$ and $(50, 60, 99)$ are assigned the same symbol $(1, 2, 3)$ (same temporal ordering) while a different symbol is assigned to $(10, 8, 14)$. Note that both symbols defined above belong to the set of possible permutations of the sets $\{1, ..., d_1\}$ and $\{1, ..., d_2\}$, i.e. in a discrete set with $d_1!$ and $d_2!$ elements respectively. This holds for any time point t, provided that the embedding dimensions are kept constant. By therefore considering multiple time points t, and deriving symbols for both the target and source at each timepoint, relative frequencies of symbols can be computed. Then, symbolic TE is defined as follows:

Definition 3.4.1 (Symbolic TE)**.** *Given two univariate time series $X_t, Y_t, t \in \mathbb{Z}$, the symbolic transfer entropy from Y to X is defined as*

$$T^S_{Y\to X} = \sum_{\widehat{x}_{t+\delta},\widehat{x}_t,\widehat{y}_t} p(\widehat{x}_{t+\delta}, \widehat{x}_t, \widehat{y}_t) \log \frac{p(\widehat{x}_{t+\delta}|\widehat{x}_t, \widehat{y}_t)}{p(\widehat{x}_{t+\delta}|\widehat{x}_t)} \tag{3.34}$$

where $\widehat{x}_{t+\delta}, \widehat{x}_t, \widehat{y}_t$ are symbols as defined in (3.32), (3.33) and δ is a time step yielding a future value of the target.

From the discussion above, despite originally assuming continuous data, we are able to estimate all probabilities involved in the definition of STE by calculating the relative frequency of all symbol combinations. After its introduction in Staniek and Lehnertz (2008) a conditional extension to STE was studied in Papana et al. (2015).

As a data transformation method, STE transforms the data to their ranks, and solely uses them to make inferences: the relation of the STE and TE resembles that of the Wilcoxon rank

sum test and t-test, the former being an application of the later on the ranks of the data. Using the ranks of the data as a proxy for their statistical dependence (an idea dating back to Spearman (1904)) essentially allows us to consider the relative magnitude ordering of each time series, and not the time series itself. Papana et al. (2015) note that this property makes it suitable for non-stationary data. However, as Wibral et al. (2013) point out, this approach implicitly assumes that all relevant information in the data lies in the ordinal relationship between values, an assumption that can potentially be misleading.

Multiple realizations of a time series

The fundamental issue with TE estimation in non-stationary data stems from the fact that the probability functions we seek to estimate are no longer time-invariant.

Wollstadt et al. (2014) (based on prior results by Gomez-Herrero et al. (2010)) indirectly propose a solution to this problem, in the case where multiple realizations of the time series involved can be obtained. This is inspired by the field of neuroscience, where multiple realizations of the same processes can be retrieved through repeated experiments. They thus propose to utilize the potential multi-trial nature of the data and estimate information theoretic quantities at each time point by searching for neighbors across all realizations and pooling them. The neighbor search occurs in consecutive overlapping time windows of fixed size. The size of each window for nearest neighbor search is thus a free parameter; for rapidly changing time series, a smaller size should be used and more realizations will be required.

While promising, this method is based on the assumption of the possibility of repeated experiments which is largely incompatible with the context of the project, and it is not pursued.

3.4.3 Stationary increments

While direct methods dealing with non-stationarity in the fully general case remain largely elusive, interesting advances exist when more assumptions are made. In a recent paper, Granero-Belinchón et al. (2019) develop a framework for the estimation of information theoretic quantities of non-stationary processes with stationary increments.

For a non-stationary time series X_t, the authors consider the differential entropy at time t. Due to non-stationarity, entropy is now time-dynamic and changes at each time point, since densities change over time as well.

Definition 3.4.2. *At time t, the differential entropy of the non-stationary time series $X_t, t \in \mathbb{Z}$ is:*

$$h_t(X) := h(X_t) = -\int p_{X_t}(x) \log p_{X_t}(x) dx \tag{3.35}$$

where p_{X_t} is the density of X_t.

Throughout the work, delay embedding vectors proved useful in studying TE. For the non-stationary time series X_t, given the embedding $X_t^{(m,\tau)}$ we write its time-dependent entropy:

$$h_t^{(m,\tau)}(X) := h(X_t^{(m,\tau)}) = -\int p_{X_t^{(m,\tau)}}(x_t^{(m,\tau)}) \log \left(p_{X_t^{(m,\tau)}}(x_t^{(m,\tau)}) \right) dx_t^{(m,\tau)} \tag{3.36}$$

Subsequently, define the time-increments of size τ for process X_t as

$$\delta_\tau X_t := X_t - L^\tau X_t = X_t - X_{t-\tau} \tag{3.37}$$

Then, consider the embedding vector

$$\widetilde{X}_t^{(m,\tau)} := (X_t, \delta_\tau X_t, \delta_\tau X_{t-\tau}, ..., \delta_\tau X_{t-(m-2)\tau}) \tag{3.38}$$

Granero-Belinchón et al. (2019) note that the following result holds:

Theorem 3.4.3. *Let $X_t, t \in \mathbb{Z}$ be a time series, and consider the following embedding vectors: $X_t^{(m,\tau)} = (X_t, X_{t-\tau}, ..., X_{t-(m-1)\tau})$ and $\widetilde{X}_t^{(m,\tau)}$ as defined in (3.38). For the differential entropy h, it holds that*

$$h(X_t^{(m,\tau)}) = h(\widetilde{X}_t^{(m,\tau)}) \tag{3.39}$$

The proof of this theorem is based on noting that $\widetilde{X}_t^{(m,\tau)}$ is a linear transformation of $X_t^{(m,\tau)}$ therefore the corollary of Theorem 8.6.4 of (Cover and Thomas, 2006, p. 254) applies.

It is then observed that for processes with stationary increments $\delta_\tau X_t$, the marginal distribution of any X_t may be time-dependent, but the marginal distribution of any increment $\delta_\tau X_t$ is not. According to the authors, the above suggest that the (problematic for estimation) time-dependence of $h(X_t^{(m,\tau)})$ mainly originates from the first component of the embedding vector X_t.

A *time-averaged* density is then defined, to be used in estimating the entropy of a non-stationary process with stationary increments. A practical framework for its estimation is also proposed.

Definition 3.4.4 (Time-averaged density)**.** *Let $X_t, t \in \mathbb{Z}$ be a non-stationary time series with stationary increments, and $[t_0, t_0 + T]$ be a time interval of length T starting at point t_0. The time-averaged probability density function of the embedding vector $X_t^{(m,\tau)}$ of X_t is*

$$\overline{p}_{T,t_0,X_t^{(m,\tau)}}(x_t^{(m,\tau)}) := \frac{1}{T}\int_{t_0}^{t_0+T} p_{X_t^{(m,\tau)}}(x_t^{(m,\tau)})dt \tag{3.40}$$

Thus, for a given embedding, this time-averaged density only depends on the starting time t_0 and on the length T of the time interval.

In practice, it is proposed to approximate the time-averaged density $\overline{p}_{T,t_0,X_t^{(m,\tau)}}$ with the histogram (see Definition 2.4.2) $\widehat{\overline{p}}_{T,t_0}$ of all data points $X_t^{(m,\tau)}$, available in that particular time interval, $t \in [t_0, t_0 + T]$.

The time-averaged density introduced will be used to estimate the entropy of an embedding $X_t^{(m,\tau)}$. The authors argue that when this is happening in the case of non-stationary processes with stationary increments, the entropy estimator will not depend on the starting time t_0 of the interval chosen, but only on its length T.

An *ersatz* entropy of $X_t^{(m,\tau)}$ is hence defined based on the time-averaged density $\overline{p}_{T,t_0,X_t^{(m,\tau)}}$

Definition 3.4.5 (Ersatz entropy)**.** *For a non-stationary time series $X = X_t, t \in \mathbb{Z}$ with stationary increments, a time interval $[t_0, t_0 + T]$ and an embedding $X_t^{(m,\tau)}$ such that $t \in [t_0, t_0 + T]$, the ersatz entropy of X is defined as*

$$\overline{h}_T^{(m,\tau)}(X) := -\int \overline{p}_{T,t_0,X_t^{(m,\tau)}}(x_t^{(m,\tau)}) \log\left[\overline{p}_{T,t_0,X_t^{(m,\tau)}}(x_t^{(m,\tau)})\right] dx_t^{(m,\tau)} \tag{3.41}$$

depending, for a given embedding, only on the interval length T.

Estimating the time-averaged density with the histogram described above, we arrive at an entropy estimator for an embedded non-stationary time series X_t with stationary increments. The ersatz entropy $\overline{h}_T^{(m,\tau)}(X)$ can be interpreted as the average uncertainty of the vector $X_t^{(m,\tau)}$ in an interval of length T.

3.5 Significance testing

As we have seen in Section 2.1.3, TE quantifies a specific conditional independence relation between the source and the target. So, in theory, provided that no directed relationship exists, the TE from a source Y to a target X would be zero. In practice, we estimate TE from data. This would generally result in a non-zero measurement, even if there is no directed relationship. Therefore, the essential question is not the value of TE itself, but whether this value is *significantly* bigger than 0.

Sketching the idea, first a null hypothesis H_0 that there is no directed relationship from Y to X is formed. To test whether the estimated TE value suggests rejecting this hypothesis in favor of the alternative hypothesis that a directed relationship exists, the distribution of the TE statistic under the null hypothesis should be known. Then, the corresponding p-value for sampling the actual TE measurement from this distribution can be computed, allowing us to reject the null hypothesis if this probability is lower than an arbitrary threshold.

So, given a dataset, we should consider how can we derive the distribution of TE values under the null hypothesis that no directed relationship from Y to X exists.

To do so, for an estimated TE value $\widehat{T}^{(k,\ell)}_{Y\to X}$ we consider the distribution of *surrogate* measurements $\widehat{T}^{(k,\ell)}_{Y^s\to X}$ under H_0. The notation Y^s represents surrogate time series for Y that are generated under H_0. These have the same statistical properties with Y, but any potential directed relationship with X is destroyed.

Except for very specific cases Barnett and Bossomaier (2012), Lizier (2014), the complexity of the TE statistic prohibits the knowledge of its surrogate measurements distribution. This has led researchers to approximate the distribution with re-sampling methods. Depending on the context, the surrogate measurement distribution can be properly approximated in the following ways:

- By shuffling (or redrawing) the $y^{(\ell)}_{t-1}$ among the set of $\{x_t, x^{(k)}_{t-1}, y^{(\ell)}_{t-1}\}$ tuples
- By rotating the time series Y (assuming stationarity)
- By swapping values for the source Y between different realizations Y_i given multiple realizations are available.

After surrogates are created, we estimate all surrogate transfer entropies $\widehat{T}^{(k,\ell)}_{Y^s\to X}$, obtaining a distribution for these values. By evaluating whether the initial TE estimate $\widehat{T}^{(k,\ell)}_{Y\to X}$ is bigger than an appropriate percentile of this distribution we are able to infer on the null hypothesis H_0.

Chapter 4

A Random Walk System

Since the estimation of transfer entropy in the non-stationary case is the research goal, two prerequisites are necessary: first, exact values for TE in a non-stationary context; and second, TE estimators to be tried. Provided these are available, it is possible to compare the estimates to the exact values of TE as well as to each other.

While TE has been a popular notion both in applications and research, studies featuring exact theoretical results for TE are scarcely found in literature, and the ones that exist pertain to stationary systems. Investigating TE in a non-stationary setting is therefore an important question, as theoretical results will enhance our theoretical understanding of TE, simultaneously enabling the possibility to evaluate TE estimators in a non-stationary context.

This chapter features a study of a specific non-stationary system including theoretical derivations and estimation of TE as well as practical insights. It starts from a stationary system that was recently studied in a relevant paper - and extrapolates results to the non-stationary case.

4.1 A stationary AR(1) system

Possibly the first study of transfer entropy featuring exact derivations was presented in Kaiser and Schreiber (2002). Simple systems were studied and solved analytically, such as binary Markov chains and Gaussian processes. Linear Gaussian systems were also analytically studied in Nichols et al. (2005) and more recently, in Novelli et al. (2019) where the authors provide exact results for TE in the context of a stationary vector autoregressive Gaussian process with linear interactions. Finally, Hahs and Pethel (2013) studied two different systems of coupled autoregressive processes. The work featured in this last paper is the main influence for the work performed in this project; the methodology of this paper is hence presented first.

Among the papers cited above, two kinds of assumptions the authors make can be discerned: a Gaussian assumption and a stationarity assumption. While the first is greatly important as the differential entropy of a (multivariate) normal distribution admits a convenient closed formula essential for analytic results, the stationarity assumption is related to its general utility in time series analysis and estimation. Starting from the stationary system Hahs and Pethel (2013) study, we will relax the second assumption, leading us to a non-stationary system to be studied.

We begin with the differential entropy of a multivariate normal random variable. For the theoretical background of multivariate normality, refer to Appendix A.

Theorem 4.1.1 (Multivariate normal entropy)**.** *Let X be a k-dimensional multivariate normal random variable, $X \sim \mathcal{N}(\mu, C)$ where μ is the mean vector and C is the covariance matrix of X. The differential entropy of X is given by the formula*

$$h(X) = \frac{1}{2} \log \left[(2\pi e)^k \det C \right] \tag{4.1}$$

Note that the entropy of such a variable only depends on its dimensionality and covariance structure; it is not influenced by the mean.

Hahs and Pethel (2013) study the following coupled system:

$$\begin{aligned} X_t &= aX_{t-1} + Z_t \\ Y_t &= bX_t + W_t \end{aligned} \tag{4.2}$$

where $|a| < 1$, $b \in \mathbb{R}_+$ and $t \in \{0, 1, 2, ...\}$. Here, Z_t, W_t are i.i.d noise processes normally distributed with zero means and variances σ_Z^2, σ_W^2 respectively. To allow for stationarity, both processes are assumed to "start" from $t = -\infty$ but are only observed from $t = 0$ onwards (see discussion in Section 2.3.2). The following independence relations are also assumed:

$$Z_t \perp\!\!\!\perp X_s, \text{ for } s < t \tag{4.3}$$
$$W_t \perp\!\!\!\perp X_s, \text{ for any } t, s \tag{4.4}$$
$$Z_t \perp\!\!\!\perp W_s, \text{ for any } t, s \tag{4.5}$$

This system represents a basic model, where X_t is a hidden time series of a quantity we would ideally want to measure, but are only able to observe through the noisy process Y_t. Indeed, at time t, the value of Y_t is approximately equal to the value of X_t up to some noise stemming from the i.i.d noise process W_t and a fixed coefficient (denoting the coupling strength) that is multiplied with the value of X_t. This is a flexible model that can be used in a variety of applications; within the context of ASML, X_t can be thought of as an underlying quantity we only observe through a sensor Y_t. We will investigate potential insights from this applied perspective later in the chapter.

4.1.1 Stationarity

Correlating the form of the system with the preliminaries presented in Chapter 2, we observe that the recursive equation defining X_t essentially defines an AR(1) process, with i.i.d noise (instead of white). We have seen that stationarity (and stability) in this case is ensured iff $|a| < 1$. Here, the structure is more complicated: a bivariate time series $U_t = (X_t, Y_t)$ is present, so stationarity must also be investigated jointly for U_t.

It is therefore clear that extensions of some time series concepts presented in Chapter 2 are required. We will focus on the bivariate case, following (Brockwell and Davis, 2010, Chapter 7).

Definition 4.1.2. *Let $U_t = (X_t, Y_t), t \in \mathbb{Z}$ be a bivariate time series.*

- *The mean vector μ and the covariance matrix Γ of U_t are defined as:*

$$\mu_t := E[U_t] = \begin{bmatrix} E[X_t] \\ E[Y_t] \end{bmatrix} \tag{4.6}$$

$$\Gamma(t+h, t) := Cov(U_{t+h}, U_t) = \begin{bmatrix} Cov(X_{t+h}, X_t) & Cov(X_{t+h}, Y_t) \\ Cov(Y_{t+h}, X_t) & Cov(Y_{t+h}, Y_t) \end{bmatrix} \tag{4.7}$$

- *U_t is said to be (weakly) stationary if μ_t and $\Gamma(t+h, t)$ are, for each h, independent of t, in which case the notation μ and $\Gamma(h)$ is used.*

These definitions allow us to extend the definition of a white noise process, that was important for AR(1) systems in the univariate case.

Definition 4.1.3. *The bivariate time series $U_t, t \in \mathbb{Z}$ is called white noise with mean 0 and covariance matrix Σ if U_t is stationary with $\mu = 0$ and*

$$\Gamma(h) = \begin{cases} \Sigma & , \text{ if } h = 0 \\ 0 & , \text{ otherwise} \end{cases} \tag{4.8}$$

From the above, a bivariate AR(1) process can be defined. Following the remarks in Section 2.3.2, the process is assumed to start at $-\infty$ but its values are only observed from $t = 0$ and on.

Definition 4.1.4. *A bivariate time series U_t, $t \in \mathbb{Z}$ is a vector autoregressive process of order 1 (VAR(1)) if for every t*

$$U_t = \Phi U_{t-1} + V_t \tag{4.9}$$

where Φ is a 2×2 real matrix and V_t is white noise with mean 0 and covariance matrix Σ.

In the corresponding discussion for the univariate AR(1) process, we showed how stability and stationarity conditions coincide in requiring $|a| < 1$ for the coefficient of the AR(1) process. The analogous result for the bivariate case pertains to the eigenvalues of matrix Φ.

Theorem 4.1.5. *As defined above, the bivariate VAR(1) process is stationary if and only if the eigenvalues of matrix Φ are less than 1 in absolute value, i.e. iff*

$$\det(I - z\Phi) \neq 0, \text{ for all } z \in \mathbb{C} \text{ such that } |z| \leq 1 \tag{4.10}$$

Now, the joint stationarity of the system (4.2) can be investigated. We first note that this model contains two i.i.d noises. Since i.i.d noise is also white noise, given that the model satisfies the requirements of Definition 4.1.4, it may be called a VAR(1) model.

First, we note that system (4.2) can be expressed in the form of Definition 4.1.4 as follows:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} a & 0 \\ ba & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} Z_t \\ bZ_t + W_t \end{bmatrix} \tag{4.11}$$

If we prove that this process is VAR(1), due to the lower triangular form of matrix Φ we would immediately obtain a condition characterizing the stationarity of this model: $|a| < 1$. The condition for stationarity of this bivariate system would therefore coincide with the univariate case. Proving the following result also proves that system (4.2) is VAR(1). The proof is straightforward and we include it in Appendix A.

Theorem 4.1.6.

$$V_t = \begin{bmatrix} Z_t \\ bZ_t + W_t \end{bmatrix} \text{ is a white noise process.} \tag{4.12}$$

In conclusion, stationarity for the model under consideration holds if and only if $|a| < 1$. Hahs and Pethel (2013) omit the above stationarity proof and directly assume this condition for a. They therefore investigate a stationary system.

4.1.2 Compensated transfer entropy

A subtle detail exists in system (4.2). Notice that, at time t, the value of Y_t depends (up to the noise W_{t-1} and coupling strength b) on the corresponding value X_t of process X *on the same timepoint* t. This means that the information flowing from process X to process Y does so with no delay; this is what we refer to as *instantaneous causation*. For simplicity, we thus assume that the sensor is able to report the value of the underlying process immediately. As it was originally defined, TE will not be able to deal with this property of this system: at time t the source embedding vector starts considering the flow of information from the source to the target at time $t - 1$. Here, the important transfer of information that TE should encapsulate is happening instantly, and the original definition of TE will miss the interaction taking place at present time.

To deal with this peculiarity, Hahs and Pethel (2013) *augment* the source embedding vector for TE in order to include the present time as well. They use the name *information transfer* instead of transfer entropy to differentiate between the two slightly different notions. So, instead of:

$$T^{(\ell,k)}_{X \to Y}(t) = I(Y_t; X^{(k)}_{t-1} | Y^{(\ell)}_{t-1}) \tag{4.13}$$

they instead consider the "information transfer"

$$IT^{(\ell,k)}_{X\to Y}(t) = I(Y_t; X_t^{(k+1)}|Y_{t-1}^{(\ell)}) \tag{4.14}$$

In a publication that preceded the work of Hahs and Pethel by a month, Faes et al. (2013b) study the same problem of instantaneous causality and TE as their main subject. Interestingly enough, the solution they propose is essentially the same: the source embedding vector is likewise augmented with the current state. They however choose the name *compensated transfer entropy* (cTE) to describe this quantity. The term "information transfer" is hence not going to be used; in what follows, cT denotes the compensated transfer entropy in a system.

Definition 4.1.7 (Compensated TE)**.** *At time t, the compensated transfer entropy from processs X_t to process Y_t is defined as follows:*

$$cT^{(\ell,k)}_{X\to Y}(t) := I(Y_t; X_t^{(k+1)}|Y_{t-1}^{(\ell)}) \tag{4.15}$$

From the conditional mutual information definition of TE and the chain rule, we derive:

$$\begin{aligned} cT^{(\ell,k)}_{X\to Y}(t) &:= I(Y_t; X_t^{(k+1)}|Y_{t-1}^{(\ell)}) && (4.16)\\ &= h(Y_t|Y_{t-1}^{(\ell)}) - h(Y_t|Y_{t-1}^{(\ell)}, X_t^{(k+1)}) && (4.17)\\ &= h(Y_t, Y_{t-1}^{(\ell)}) - h(Y_{t-1}^{(\ell)}) - h(Y_t, Y_{t-1}^{(\ell)}, X_t^{(k+1)}) + h(Y_{t-1}^{(\ell)}, X_t^{(k+1)}) && (4.18)\end{aligned}$$

4.1.3 Approach and limitations

Hahs and Pethel compute exact cTE values over arbitrary embeddings as in (4.15). To do so, they start from an analytic computation for $cT_{X\to Y}$ over a single time lag $t-1 \to t$, i.e. for $k = \ell = 1$ in (4.15):

$$\begin{aligned} cT^{(1,1)}_{X\to Y}(t) &= I(Y_t; X_t^{(2)}|Y_{t-1}^{(1)}) && (4.19)\\ &= h(Y_t, Y_{t-1}^{(1)}) - h(Y_{t-1}^{(1)}) - h(Y_t, Y_{t-1}^{(1)}, X_t^{(2)}) + h(Y_{t-1}^{(1)}, X_t^{(2)}) && (4.20)\\ &= h(Y_t, Y_{t-1}) - h(Y_{t-1}) - h(Y_t, Y_{t-1}, X_t, X_{t-1}) + h(Y_{t-1}, X_t, X_{t-1}) && (4.21)\end{aligned}$$

This cTE is fully specified by the covariance matrix of the random vector $[X_t, Y_t, X_{t-1}, Y_{t-1}]^T$. There are three reasons for this:

1. Due to the assumptions made, the random vector $[X_t, Y_t, X_{t-1}, Y_{t-1}]^T$ is multivariate normal
2. All random variables listed in (4.21) are elements ("subsets") of the aforementioned random vector hence still (multivariate) normal (see relevant property in Appendix A)
3. The entropy of a (multivariate) normal random variable (4.1) depends on the covariance matrix and its dimension only

After deriving the value of the above cTE, the important step in the approach of Hahs and Pethel is based on the fact that, because the system is stationary, the aforementioned covariance matrix (and therefore the cTE too) does not depend on the time t. This allows the utilization of this covariance matrix as a building block for more general covariance matrices enabling the calculation of cTE over arbitrary embedding vectors. This is achieved through a recursive approach.

However, for a non-stationary system, this is no longer possible - since the covariance matrix will be changing over time. Moreover, while the second and third points listed above are clear, the fact that this specific random vector is multivariate normal has to be proven.

In the next section, a non-stationary system is introduced and studied; it is a limiting case of model (4.2) on the boundary of its parameter space. As a similar recursive approach is infeasible, we resort to explicit derivations of the quantities needed. We also prove the normality of the random vector mentioned before. This is omitted in Hahs and Pethel (2013). Then, a simple extension of this non-stationary system is studied, and its implications in the findings are investigated.

4.2 Non-stationary extension

Recall the VAR(1) model that was studied in previous section:

$$\begin{aligned} X_t &= aX_{t-1} + Z_t \\ Y_t &= bX_t + W_t \end{aligned} \tag{4.22}$$

Based on VAR(1) theory, stationarity of this system was found to be equivalent with the condition $|a| < 1$. We will now consider the case where $a = 1$. The system is then non-stationary and it takes the form:

$$\begin{aligned} X_t &= X_{t-1} + Z_t \\ Y_t &= bX_t + W_t \end{aligned} \tag{4.23}$$

where $t \in \{0, 1, 2, ...\}$, $X_0 = Y_0 = 0$, and processes Z_t, W_t are i.i.d normal noises with mean 0 and variances σ_Z^2, σ_W^2 respectively. Independence assumptions are as before:

$$Z_t \perp\!\!\!\perp X_s, \text{ for } s < t \tag{4.24}$$
$$W_t \perp\!\!\!\perp X_s, \text{ for any } t, s \tag{4.25}$$
$$Z_t \perp\!\!\!\perp W_s, \text{ for any } t, s \tag{4.26}$$

We start by investigating the distribution of the random variables comprising this system, and then examine the stationarity of its increments.

4.2.1 Distribution

Rewriting the equation for X_t we obtain

$$X_t = X_{t-1} + Z_t \tag{4.27}$$
$$= X_{t-2} + Z_{t-1} + Z_t \tag{4.28}$$
$$= X_{t-3} + Z_{t-2} + Z_{t-1} + Z_t \tag{4.29}$$
$$= ... \tag{4.30}$$
$$= X_0 + Z_1 + ... + Z_t \tag{4.31}$$

Therefore,

$$X_t = \sum_{k=1}^{t} Z_k \tag{4.32}$$

Hence, the process X_t is a random walk, as defined in Chapter 2. Since the process Z_t is i.i.d normal noise we are able to calculate the distribution of X_t. For fixed t, we have

$$Z_t \sim \mathcal{N}(0, \sigma_Z^2) \tag{4.33}$$

Also, $\{Z_1, ..., Z_t\}$ are independent, therefore (sum of independent normals)

$$X_t = \sum_{k=1}^{t} Z_k \sim \mathcal{N}(0, t\sigma_Z^2) \tag{4.34}$$

As noted in the corresponding section, random walks are non-stationary, with an autocovariance function $\gamma(t + h, t) = t\sigma_Z^2$. Utilizing these results we may also obtain the distribution of Y_t:

$$X_t \sim \mathcal{N}(0, t\sigma_Z^2) \implies bX_t \sim \mathcal{N}(0, b^2 t\sigma_Z^2) \tag{4.35}$$

Due to the independence of W_t and X_t we then get

$$bX_t + W_t = Y_t \sim \mathcal{N}(0, \sigma_W^2 + b^2 t\sigma_Z^2) \tag{4.36}$$

To conclude, both processes X_t and Y_t of the model are, for a fixed time t, normal random variables. Their marginal non-stationarity can also be observed by the dependence of their distribution on time. Then, we define the joint process:

$$U_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} \tag{4.37}$$

At time t, the bivariate random variable U_t has mean vector

$$\mu_t = \begin{bmatrix} E[X_t] \\ E[Y_t] \end{bmatrix} \tag{4.38}$$

and covariance matrix

$$\Sigma_t = \begin{bmatrix} Var(X_t) & Cov(X_t, Y_t) \\ Cov(Y_t, X_t) & Var(Y_t) \end{bmatrix} \tag{4.39}$$

Both quantities can be easily computed explicitly. For the general case, we are interested in an arbitrary embedding of the joint process U_t:

$$U_t^{(d)} = (U_t, U_{t-1}, ..., U_{t-(d-1)}) \tag{4.40}$$

So, at time t, the multivariate random variable we will consider is

$$U_t^{(d)} = \begin{bmatrix} X_t \\ Y_t \\ X_{t-1} \\ Y_{t-1} \\ \vdots \\ X_{t-(d-1)} \\ Y_{t-(d-1)} \end{bmatrix} \tag{4.41}$$

This multivariate random variable has a mean vector μ_t and covariance matrix Σ_t that are similarly defined as in the bivariate case. We already saw that all elements of this random vector are marginally (univariate) normal. We will now prove that the embedding $U_t^{(d)}$ follows a multivariate normal distribution.

Theorem 4.2.1. *The random vector $U_t^{(d)}$ follows a multivariate normal distribution.*

Proof. The proof revolves around noting that $U_t^{(d)}$ can be decomposed as a product of a real matrix and a normally distributed random vector. Then, because multivariate normality is preserved under affine transformations (see Appendix A) the embedding $U_t^{(d)}$ will also be multivariate normal with a specific mean and covariance matrix that can be computed from the decomposition. We write:

$$\begin{bmatrix} X_t \\ Y_t \\ X_{t-1} \\ Y_{t-1} \\ \vdots \\ X_{t-(d-1)} \\ Y_{t-(d-1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 & \dots & 1 & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ b & b & \dots & b & \dots & b & b & 1 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 1 & \dots & 1 & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & b & \dots & b & \dots & b & b & 0 & 1 & \dots & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 1 & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & b & \dots & b & b & 0 & 0 & \dots & 1 & \dots & 0 & 0 \end{bmatrix} \begin{bmatrix} Z_t \\ Z_{t-1} \\ \vdots \\ Z_{t-(d-1)} \\ \vdots \\ Z_2 \\ Z_1 \\ W_t \\ W_{t-1} \\ \vdots \\ W_{t-(d-1)} \\ \vdots \\ W_2 \\ W_1 \end{bmatrix}$$

Therefore we have decomposed the embedding as:

$$U_t^{(d)} = A \cdot L \tag{4.42}$$

where A is a $2d \times 2t$ real matrix and L is a $2t$-dimensional random vector. Due to the i.i.d assumptions of the model, the random vector L follows a multivariate normal distribution:

$$L \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_Z^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_Z^2 & & & \\ & & & \sigma_W^2 & & \\ & & & & \ddots & \\ & & & & & \sigma_W^2 \end{bmatrix} \right) \tag{4.43}$$

Hence, $U_t^{(d)} = A \cdot L$ has the following multivariate normal distribution:

$$U_t^{(d)} \sim \mathcal{N}(0, A\Sigma A^T) \tag{4.44}$$

where 0 is the d-dimensional zero vector and Σ is the covariance matrix of L listed above. □

The covariance matrix $A\Sigma A^T$ is computed:

$$\begin{bmatrix} t\sigma_Z^2 & bt\sigma_Z^2 & (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & \dots & (t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 \\ bt\sigma_Z^2 & b^2t\sigma_Z^2 + \sigma_W^2 & b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2 & \dots & b(t-(d-1))\sigma_Z^2 & b^2(t-(d-1))\sigma_Z^2 \\ (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & \dots & (t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 \\ b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2 + \sigma_W^2 & \dots & b(t-(d-1))\sigma_Z^2 & b^2(t-(d-1))\sigma_Z^2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ (t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 & (t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 & \dots & (t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 \\ b(t-(d-1))\sigma_Z^2 & b^2(t-(d-1))\sigma_Z^2 & b(t-(d-1))\sigma_Z^2 & b^2(t-(d-1))\sigma_Z^2 & \dots & b(t-(d-1))\sigma_Z^2 & b^2(t-(d-1))\sigma_Z^2 + \sigma_W^2 \end{bmatrix} \tag{4.45}$$

Given the model parameters $b, \sigma_Z^2, \sigma_W^2$, the embedding dimension d numerically specifies the covariance matrix $A\Sigma A^T$ of the multivariate normal variable $U_t^{(d)}$. Recalling that the differential entropy of such a variable only depends on its dimension and covariance matrix, we note that the embedding dimension d also numerically specifies each entropy term featured in (4.21) and therefore the cTE too. A concrete example is studied in Section 4.3.1.

4.2.2 Stationarity of increments

From the discussion of stationarity for the initial system we already saw that choosing $a = 1$ implies the non-stationarity of the bivariate system (4.23). Minding the estimator discussed in Section 3.4.3 we then focus on the stationarity of the increments of the system - recall Definitions 2.3.10 and 4.1.2.

We wish to prove the increment stationarity of the marginal processes X_t, Y_t and joint process U_t. In Section 4.3 we will be focusing on the 2-dimensional embedding of the joint process $U_t^{(2)} = [X_t, Y_t, X_{t-1}, Y_{t-1}]^T$ so stationarity of increments will be proven here for this 4-dimensional time series. This implies that any subvector of this embedding also has stationary increments, including, among others, the marginal and joint processes X_t, Y_t, U_t. We shortly elaborate on this statement below. The increments of $U_t^{(2)}$ are formed:

$$(I-B)U_t^{(2)} = U_t^{(2)} - U_{t-1}^{(2)} = \begin{bmatrix} \Delta X_t \\ \Delta Y_t \\ \Delta X_{t-1} \\ \Delta Y_{t-1} \end{bmatrix} = \begin{bmatrix} X_t - X_{t-1} \\ Y_t - Y_{t-1} \\ X_{t-1} - X_{t-2} \\ Y_{t-1} - Y_{t-2} \end{bmatrix} = \begin{bmatrix} Z_t \\ bZ_t + W_t - W_{t-1} \\ Z_{t-1} \\ bZ_{t-1} + W_{t-1} - W_{t-2} \end{bmatrix} \tag{4.46}$$

Stationarity of this multivariate time series is assessed as before. First, the mean vector is

$$\mu = \begin{bmatrix} E[Z_t] \\ E[bZ_t + W_t - W_{t-1}] \\ E[Z_{t-1}] \\ E[bZ_{t-1} + W_{t-1} - W_{t-2}] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4.47}$$

and is therefore trivially independent of time t. The same should hold for the 4×4 covariance matrix $\Gamma(t+h, t)$. For brevity, we use the first difference operator Δ and refer to (4.46) for the actual random variables under consideration:

$$\begin{bmatrix} Cov(\Delta X_{t+h}, \Delta X_t) & Cov(\Delta X_{t+h}, \Delta Y_t) & Cov(\Delta X_{t+h}, \Delta X_{t-1}) & Cov(\Delta X_{t+h}, \Delta Y_{t-1}) \\ Cov(\Delta Y_{t+h}, \Delta X_t) & Cov(\Delta Y_{t+h}, \Delta Y_t) & Cov(\Delta Y_{t+h}, \Delta X_{t-1}) & Cov(\Delta Y_{t+h}, \Delta Y_{t-1}) \\ Cov(\Delta X_{t-1+h}, \Delta X_t) & Cov(\Delta X_{t-1+h}, \Delta Y_t) & Cov(\Delta X_{t-1+h}, \Delta X_{t-1}) & Cov(\Delta X_{t-1+h}, \Delta Y_{t-1}) \\ Cov(\Delta Y_{t-1+h}, \Delta X_t) & Cov(\Delta Y_{t-1+h}, \Delta Y_t) & Cov(\Delta Y_{t-1+h}, \Delta X_{t-1}) & Cov(\Delta Y_{t-1+h}, \Delta Y_{t-1}) \end{bmatrix} \tag{4.48}$$

Using the bilinear property of the covariance and recalling the independence assumptions of the model, we note that all covariances comprising the covariance matrix are, for an arbitrary h, either 0 or a function of the variances σ_Z^2, σ_W^2 that does not depend on time. As an example, we calculate:

$$Cov(\Delta X_{t-1+h}, \Delta Y_t) = Cov(Z_{t+h-1}, bZ_t + W_t - W_{t-1}) = bCov(Z_{t+h-1}, Z_t) = \begin{cases} b\sigma_Z^2 & \text{if } h = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.49}$$

We thus conclude that the time series $(I-B)U_t^{(2)}$ is stationary, i.e. the 2-dimensional embedding $U_t^{(2)}$ of the joint process U_t has stationary increments. Observing that the mean vector and covariance matrix of any subvector of $U_t^{(2)} = [X_t, Y_t, X_{t-1}, Y_{t-1}]^T$ is a subvector of μ and a submatrix of $\Gamma(t+h, t)$ respectively, we note that they will also be independent of time. This proves that any subvector of $U_t^{(2)}$ (including the marginal and joint time series X_t, Y_t, U_t) also has stationary increments. This remark will be useful in Section 4.3.3.

Remark. *Focusing on the bivariate time series U_t, we essentially showed that it can be transformed from non-stationary to stationary through one application of the differencing operator Δ. In time series literature, this is referred to as U_t being an integrated time series of order 1. Integration is related to the similar concept of co-integration that was introduced in a highly influential paper*

by Engle and Granger (1987). In (Brockwell and Davis, 2010, Section 7.7) the same system we study is investigated from a time series analysis perspective, featuring comments on co-integration and an alternative proof for stationarity of increments.

4.2.3 Adding a deterministic drift

An immediate extension of the results obtained so far is provided by considering the case of a random walk with a drift.

Definition 4.2.2. *A time series S_t, $t \in \{0, 1, ...\}$ with $S_0 = 0$ is called a random walk with a drift μ if*

$$S_t = \mu + S_{t-1} + Z_t, \text{ for } t = 1, 2, .. \tag{4.50}$$

where $\mu \in \mathbb{R}$ is constant and Z_t is i.i.d noise.

The random walk with a drift admits a similar form to the random walk by successive substitution:

$$S_t = \mu + S_{t-1} + Z_t = 2\mu + S_{t-2} + Z_t + Z_{t-1} = ... = t\mu + \sum_{k=1}^{t} Z_k \tag{4.51}$$

The sole difference with the regular random walk is thus the linear term $t\mu$. Because of this term, for a given t, the mean of the random variable S_t is no longer 0:

$$E[S_t] = E[t\mu] + E\big[\sum_{k=1}^{t} Z_k\big] = t\mu + \sum_{k=1}^{t} E[Z_k] = t\mu \tag{4.52}$$

but since the new term is deterministic the autocovariance is not affected:

$$\gamma_s(t+h, t) = Cov(S_{t+h}, S_t) = Cov\big((t+h)\mu + S_t + Z_{t+1} + ... + Z_{t+h}, S_t\big) = Cov(S_t, S_t) = t\sigma^2 \tag{4.53}$$

This remark already indicates that the results for TE may not change. As we already discussed, TE depends on joint entropies that for a multivariate normal variable depend on its dimension and covariance matrix only. We elaborate on this statement here, following the same ideas as before. We consider the extended model:

$$X_t = \mu + X_{t-1} + Z_t \tag{4.54}$$

$$= t\mu + \sum_{k=1}^{t} Z_k \tag{4.55}$$

$$Y_t = bX_t + W_t \tag{4.56}$$

$$= bt\mu + b\sum_{k=1}^{t} Z_k + W_t \tag{4.57}$$

The assumptions regarding Z_t, W_t and independence relations are as before. For fixed t we infer:

$$X_t \sim \mathcal{N}(t\mu, t\sigma^2) \tag{4.58}$$

$$Y_t \sim \mathcal{N}(bt\mu, b^2 t\sigma_Z^2 + \sigma_W^2) \tag{4.59}$$

Then, the distribution of an embedding $U_t^{(d)}$ of the bivariate process $U_t = [X_t, Y_t]^T$ can be computed. We note that the only difference of the current model with the previous one is the existence of the deterministic terms $t\mu$ and $bt\mu$ in the equations of X_t and Y_t respectively. Hence, we observe

that the decomposition $A \cdot L$, where A and L are as in (4.42) can be trivially extended to yield $U_t^{(d)}$ by adding the following deterministic vector to it:

$$D = \begin{bmatrix} t\mu \\ bt\mu \\ (t-1)\mu \\ b(t-1)\mu \\ \vdots \\ \big(t-(d-1)\big)\mu \\ b\big(t-(d-1)\big)\mu \end{bmatrix} \tag{4.60}$$

That is, for fixed t, the new decomposition is

$$U_t^{(d)} = A \cdot L + D \tag{4.61}$$

where A, L are as in (4.42) and D is as defined above. As an affine transformation of the multivariate normal vector L (that has the same covariance matrix Σ as before), the distribution of $U_t^{(d)}$ is multivariate normal with parameters:

$$U_t^{(d)} \sim \mathcal{N}(D, A\Sigma A^T) \tag{4.62}$$

Since the covariance matrix of the current embedding vector remained the same, we infer that TE will also stay the same.

4.3 Transfer entropy insights

The theoretical analysis has now concluded, and the attention is shifted to practical aspects. In this section, an explicit formula for the cTE is derived, and its asymptotic behavior is examined. Exact results are subsequently obtained and insights on the behavior of the cTE are retrieved through sensitivity analysis. Finally, cTE is estimated, and the estimates are compared to the theoretical values. Various visualizations are included in the section to aid understanding and illuminate insights.

4.3.1 Explicit formula and asymptotic behavior

Building upon the results derived so far, a concrete cTE calculation is now presented. An exact formula for cTE is obtained, and its asymptotic behavior is examined. The following cTE is considered:

$$cT_{X\to Y}^{(1,1)}(t) = h(Y_t, Y_{t-1}) - h(Y_{t-1}) - h(Y_t, Y_{t-1}, X_t, X_{t-1}) + h(Y_{t-1}, X_t, X_{t-1}) \tag{4.63}$$

The corresponding embedding of the joint process is

$$U_t^{(2)} = \begin{bmatrix} X_t \\ Y_t \\ X_{t-1} \\ Y_{t-1} \end{bmatrix} \tag{4.64}$$

$U_t^{(2)}$ is normal with covariance matrix $A\Sigma A^T$ where A is the following $4 \times (t+2)$ matrix

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 & 0 & 0 \\ b & b & \dots & b & b & 1 & 0 \\ 0 & 1 & \dots & 1 & 1 & 0 & 0 \\ 0 & b & \dots & b & b & 0 & 1 \end{bmatrix} \tag{4.65}$$

and Σ is the following $(t+2)\times(t+2)$ diagonal matrix:

$$\Sigma = \begin{bmatrix} \sigma_Z^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_Z^2 & & & \\ & & & \sigma_W^2 & & \\ & & & & \sigma_W^2 & \end{bmatrix} \tag{4.66}$$

The end result $A\Sigma A^T$ is the 4×4 matrix

$$M := A\Sigma A^T = \begin{bmatrix} t\sigma_Z^2 & bt\sigma_Z^2 & (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 \\ bt\sigma_Z^2 & b^2t\sigma_Z^2+\sigma_W^2 & b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2 \\ (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & (t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 \\ b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2 & b(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2+\sigma_W^2 \end{bmatrix} \tag{4.67}$$

whose (i,j) element is the covariance of the i^{th} element of $U_t^{(2)}$ with its j^{th} element. The main diagonal contains the variances of $X_t, X_{t-1}, Y_t, Y_{t-1}$ which we could already compute from the autocovariance functions of each marginal process.

The matrix $A\Sigma A^T$ illustrates the dependence of the covariance structure of the model (and therefore of its cTE) to time. To calculate the cTE we introduce some notation.

Definition 4.3.1. *Let A be a $m\times n$ matrix and $k\leq n,m$. Given a set of distinct indices $i_1,...,i_k$ we will denote the matrix that only consists of the $(i_1,...,i_k)$ columns and rows of A as $A_{(i_1,...,i_k)}$.*

For instance, we are interested in computing the differential entropy $h(Y_t, Y_{t-1})$. This is a bivariate normal random variable and its covariance matrix is retrieved from (4.67) by considering the elements on each intersection of the second and fourth rows and columns of matrix M, i.e. by considering the 2×2 matrix

$$M_{(2,4)} = \begin{bmatrix} b^2t\sigma_Z^2+\sigma_W^2 & b^2(t-1)\sigma_Z^2 \\ b^2(t-1)\sigma_Z^2 & b^2(t-1)\sigma_Z^2+\sigma_W^2 \end{bmatrix} \tag{4.68}$$

Each entropy term of (4.63) is then calculated considering formula (4.1) thus yielding the cTE:

$$\begin{aligned} cT_{X\to Y}^{(1,1)}(t) &= h(Y_t,Y_{t-1}) - h(Y_{t-1}) - h(X_t,Y_t,X_{t-1},Y_{t-1}) + h(X_t,X_{t-1},Y_{t-1}) && (4.69)\\ &= \frac{1}{2}\log\left[(2\pi e)^2\det M_{(2,4)}\right] - \frac{1}{2}\log\left[(2\pi e)^1\det M_{(4)}\right] && (4.70)\\ &\qquad - \frac{1}{2}\log\left[(2\pi e)^4\det M\right] + \frac{1}{2}\log\left[(2\pi e)^3\det M_{(1,3,4)}\right] && (4.71)\\ &= \frac{1}{2}\log\left[2\pi e\frac{\det M_{(2,4)}}{\det M_{(4)}}\right] - \frac{1}{2}\log\left[2\pi e\frac{\det M}{\det M_{(1,3,4)}}\right] && (4.72)\\ &= \frac{1}{2}\log\left[\frac{\det M_{(2,4)}\det M_{(1,3,4)}}{\det M_{(4)}\det M}\right] && (4.73) \end{aligned}$$

Note that since the "matrix" $M_{(4)}$ refers to the single element included in the intersection of the fourth row and the fourth column, $M_{(4)}$ is actually the number $b^2(t-1)\sigma_Z^2+\sigma_W^2$ (the variance of the random variable Y_{t-1}), and the notation $\det M_{(4)}$ is superfluous; we however keep it for homogeneity reasons. Specifying the three parameters included in matrix M we then use formula (4.74) to derive the $cT_{X\to Y}^{(1,1)}$ as a function of time t. We may subsequently compute the determinants and substitute them in the above formula to derive a more explicit formula:

$$cT_{X\to Y}^{(1,1)}(t) = \frac{1}{2}\log\left[\frac{b^4\sigma_Z^4(t-1)+b^2\sigma_W^2\sigma_Z^2(2t-1)+\sigma_W^4}{b^2\sigma_Z^2\sigma_W^2(t-1)+\sigma_W^4}\right] \tag{4.74}$$

The limit to infinity is also calculated:

$$\lim_{t\to+\infty} cT^{(1,1)}_{X\to Y}(t) = \lim_{t\to+\infty} \frac{1}{2}\log\left[\frac{b^4\sigma_Z^4(t-1)+b^2\sigma_W^2\sigma_Z^2(2t-1)+\sigma_W^4}{b^2\sigma_Z^2\sigma_W^2(t-1)+\sigma_W^4}\right] \tag{4.75}$$

$$= \lim_{t\to+\infty} \frac{1}{2}\log\left[\frac{b^4\sigma_Z^4+b^2\sigma_W^2\sigma_Z^2\frac{2t-1}{t-1}+\frac{\sigma_W^4}{t-1}}{b^2\sigma_Z^2\sigma_W^2+\frac{\sigma_W^4}{t-1}}\right] \tag{4.76}$$

$$= \frac{1}{2}\log\left[\frac{b^4\sigma_Z^4+2b^2\sigma_W^2\sigma_Z^2}{b^2\sigma_Z^2\sigma_W^2}\right] \tag{4.77}$$

$$= \frac{1}{2}\log\left[b^2\frac{\sigma_Z^2}{\sigma_W^2}+2\right] \tag{4.78}$$

Thus, assuming that $b, \sigma_Z^2, \sigma_W^2 \in \mathbb{R}_+$, the cTE of this system converges to a constant. The (asymptotic) dependence of the information transfer magnitude on the fixed system parameters $b, \sigma_Z^2, \sigma_W^2$ is given by (4.78). Practical insights from this result are discussed in the next section.

4.3.2 Exact results and sensitivity analysis

In this section, we select the following parameter values: $\sigma_Z^2 = 0.5, \sigma_W^2 = 1$ and $b = 0.8$. We are thus able to explicitly calculate the cTE of the embedding $U_t^{(2)}$ utilizing formula (4.74). The cTE is visualized as a function of time in Figure 4.1.

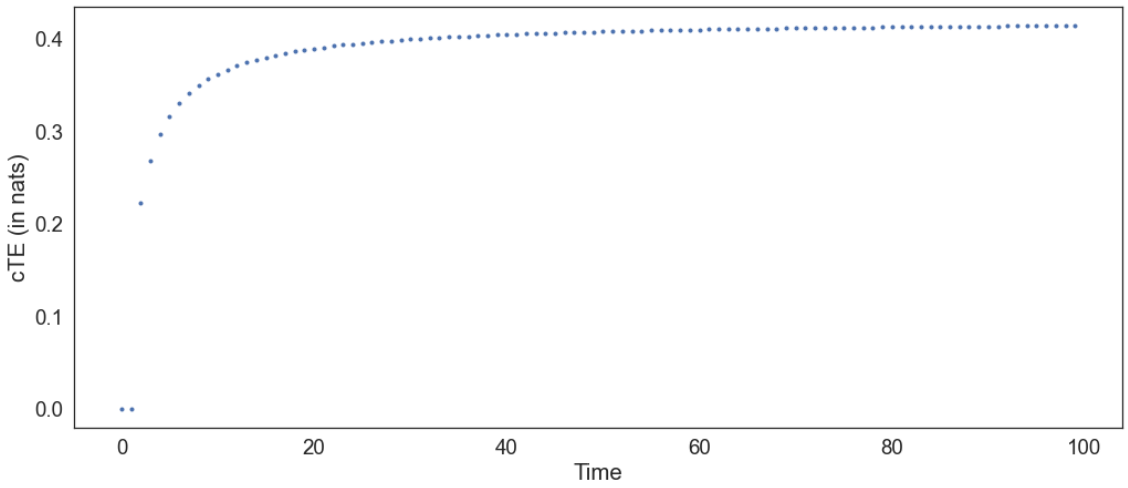

Figure 4.1: The compensated transfer entropy theoretical value for model (4.23) with $\sigma_Z^2 = 0.5, \sigma_W^2 = 1$ and $b = 0.8$. The cTE is a function of time given by formula (4.74).

Referring to Figure 4.1, note that the convergence of the cTE is clear. The limit is conveniently computed from (4.78):

$$\lim_{t\to+\infty} cT^{(1,1)}_{X\to Y}(t) = \frac{1}{2}\log\left[b^2\frac{\sigma_Z^2}{\sigma_W^2}+2\right] = \frac{1}{2}\log\left[0.8^2\frac{0.5}{1}+2\right] = \frac{1}{2}\log(2.32) = 0.421 \quad \text{nats.} \tag{4.79}$$

Sensitivity Analysis

Recall that X_t can be thought of as a hidden time series of a quantity we wish to measure, while Y_t is the observation of X_t we are able to get through a sensor. Then, σ_Z^2 is related to the variance of

X_t, σ_W^2 is the variance of the noise of Y_t, i.e. how noisy the sensor Y_t is, and b is a fixed coefficient of the values of X_t in the equation of Y_t.

The cTE from X to Y is the information transfer (in nats) from the hidden process to its sensor. Intuitively, this should be kept high; it may therefore serve as an evaluation metric for sensor Y_t. For this particular system, the cTE converges fast as time progresses, and sensitivity analysis is thus performed on its limit.

We begin with investigating the effect of sensor noise on information transfer. We fix $\sigma_Z^2 = b = 1$ and vary σ_W^2 on the interval $(0, 10)$. The cTE is plotted as a function of σ_W^2 in Figure 4.2.

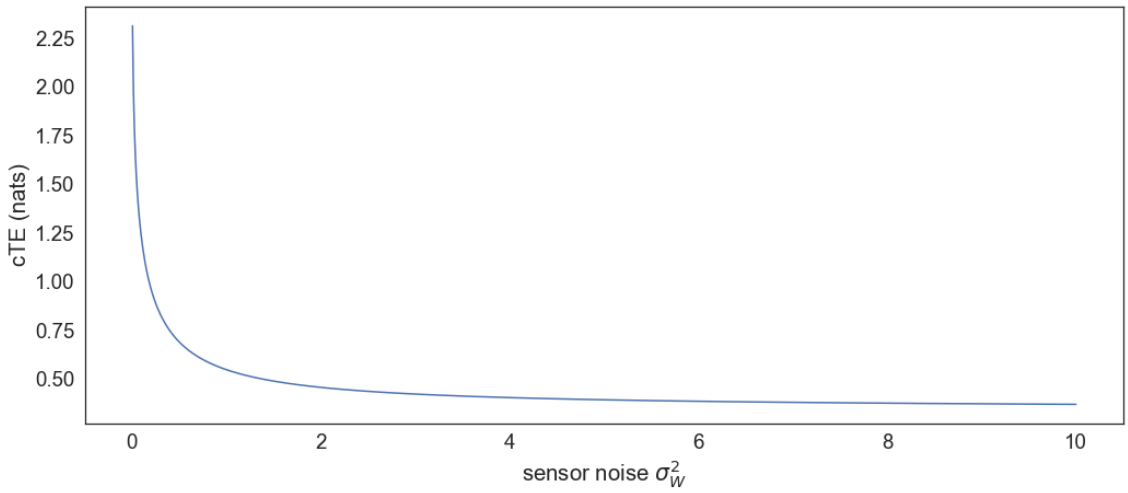

Figure 4.2: The effect of an increasing sensor noise on cTE. As the sensor becomes more noisy, the flow of information deteriorates.

A rapid descent is observed. When the system parameters $b, \sigma_W^2, \sigma_Z^2$ are positive real numbers, the asymptotic transfer of information quickly drops as the sensor noise increases (while staying finite). After having taken $t \to +\infty$ in (4.74), caution is required if limits of other system parameters are considered - the order of taking the limits is important and they generally cannot be interchanged.

We now study the effect of b on the cTE. Later in the report, we will be referring to such coefficients as the coupling strength of a causal interaction. Variances σ_Z^2, σ_W^2 are fixed to 1 and coefficient b is varied in $(0, 10)$. The corresponding graph is included in Figure 4.3.

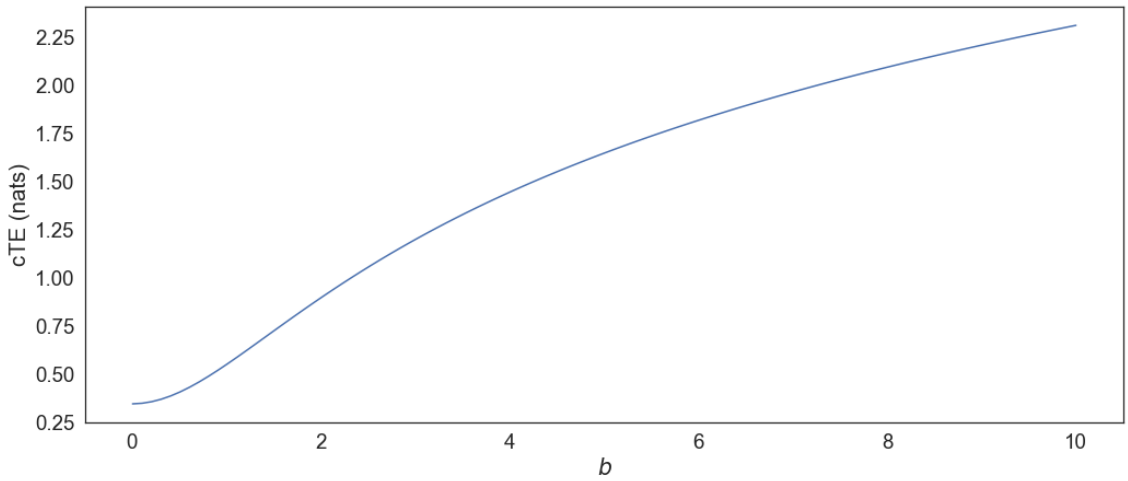

Figure 4.3: Increasing the coupling coefficient b, the information flow to the sensor increases.

Since the coefficient b appears as a square in the cTE formula, for very small values (where the sensor is reducing the correct hidden process value) the flow of information is small, however as b transitions to bigger values (where the sensor is magnifying the correct hidden process value and thus minimizing the effect of its own noise on its value) cTE is increased.

In the following, we will concentrate on the case where the coupling coefficient $b = 1$, and the system is fully determined by the variances σ_Z^2, σ_W^2 of the noise terms. From formula (4.78) it is then clear that the deciding quantity for the limit of information flow in this system is actually the ratio of variances $\frac{\sigma_Z^2}{\sigma_W^2}$. Varying this ratio on the interval $(0, 10)$, Figure 4.4 shows that the increase in information transfer from X_t to Y_t is indeed logarithmic.

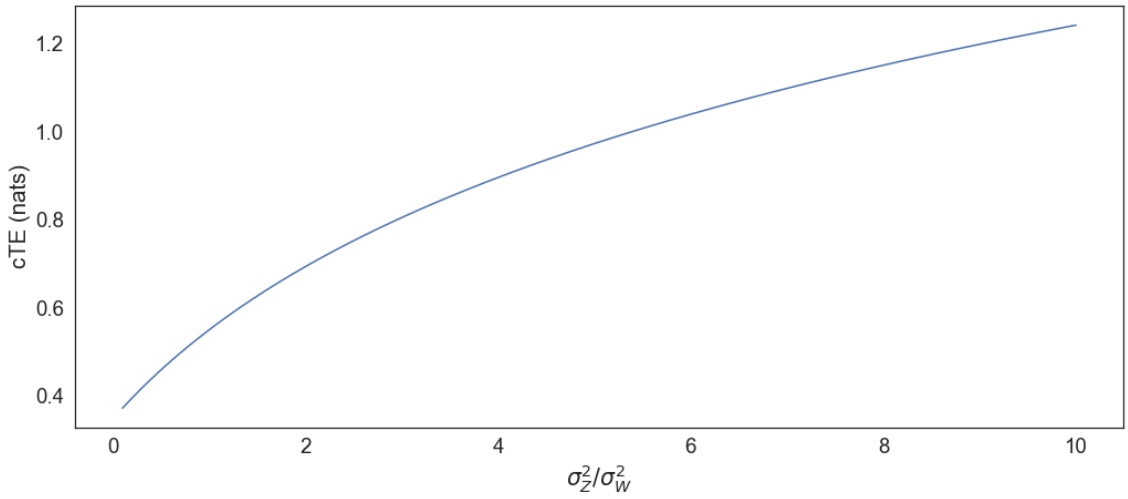

Figure 4.4: Increasing the variances ratio implies a logarithmic increase in information flow.

We conclude this analysis by varying both variances σ_Z^2, σ_W^2 at the same time, and visualizing the cTE for each combination in $(0, 10) \times (0, 10)$. This 3D graph is shown in Figure 4.5.

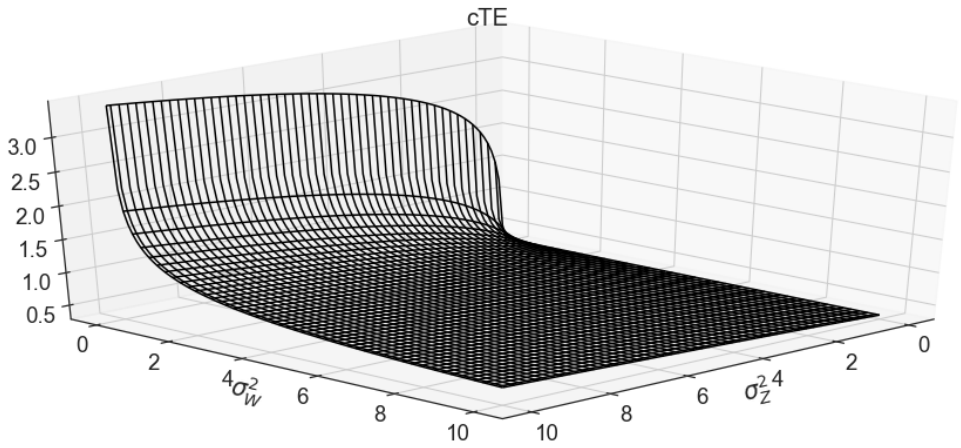

Figure 4.5: Varying both variances on the same interval as before, we obtain a 3-dimensional graph displaying how cTE changes. Observe the sharp drop in information transfer as the sensor noise increases, which is even sharper when the hidden process variability is small (smaller than the sensor noise). After the initial drop, cTE resembles a slightly inclined plane.

4.3.3 Estimator performance

A specific realization of the model is now generated. It is visualized in Figure 4.6.

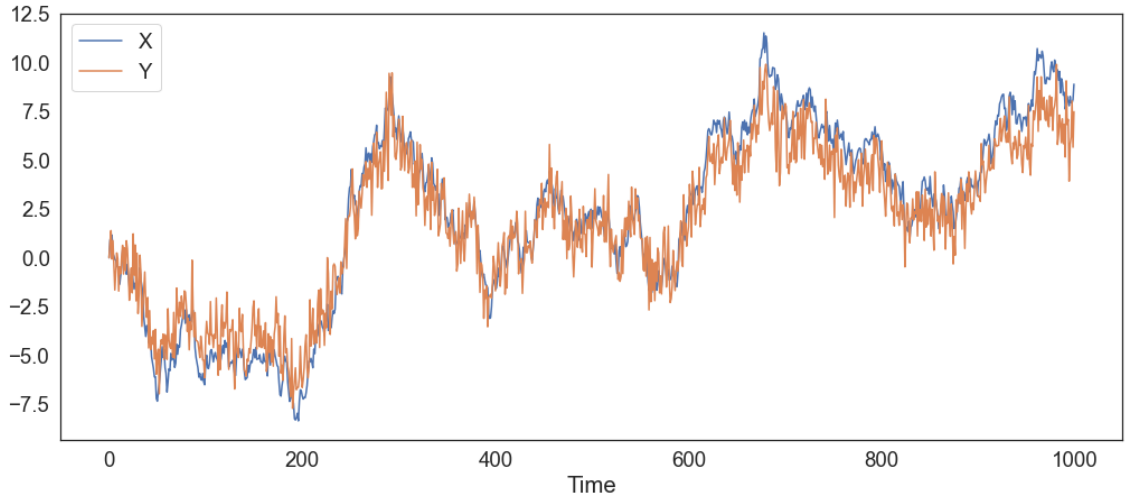

Figure 4.6: A realization of model (4.23) with the same parameters as in Figure 4.1. Process Y_t is a noisy observation of X_t.

We will use this realization of the system to estimate the cTE whose exact values and limit are already known. We are then able to assess the performance of the estimator we choose for this system: the estimator presented in Section 3.4.3 will be implemented for each entropy term in the expression for cTE (4.63). As shown in Section 4.2.2, each entropy term refers to a non-stationary time series with stationary increments, so this estimator is a suitable fit.

As we saw in Figure 4.1, the exact cTE is a function of time, albeit quickly convergent to a constant value. Therefore, creating a large time window to estimate the cTE on, we will first compare the single estimate obtained with the limiting value of the cTE.

We select the time window $[150, 450]$. Referring to Figure 4.6 a significant portion of this window contains a non-stationary pattern in the form of drifts.

In Section 3.4.3 it is proposed to estimate the time-averaged density (3.40) of the window $[150, 450]$ with the histogram of all data points available in that time interval: the entropy estimate of that time window is then the entropy of that histogram. For the purposes of the project, the following multivariate density estimator was used instead Lizier (2014):

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \Theta\left(||x - X_i||_\infty - h\right) \tag{4.80}$$

Here, Θ is the step kernel $\Theta(x > 0) = 0$, $\Theta(x \leq 0) = 1$, and $||\cdot||_\infty$ is the maximum norm. We note that such an estimate might require further calibration with respect to its bias; potential bias correction for estimates is however not considered in this study.

Each cTE term $h(Y_t, Y_{t-1}), h(Y_{t-1}), h(Y_t, Y_{t-1}, X_t, X_{t-1}), h(Y_{t-1}, X_t, X_{t-1})$ is thus separately estimated using this estimator in the time window $[150, 450]$, i.e. with $T = 300$. All estimates are in nats, to match the cTE theoretical values. We obtain:

$$\widehat{h}(Y_t, Y_{t-1}) = 4.448 \tag{4.81}$$

$$\widehat{h}(Y_{t-1}) = 2.709 \tag{4.82}$$

$$\widehat{h}(Y_t, Y_{t-1}, X_t, X_{t-1}) = 6.637 \tag{4.83}$$

$$\widehat{h}(Y_{t-1}, X_t, X_{t-1}) = 5.303 \tag{4.84}$$

The cTE estimate for this particular time window therefore is:

$$\widehat{cT}_{X\to Y}^{(1,1)(150,450)} = 4.448 - 2.709 - 6.637 + 5.303 = 0.405 \tag{4.85}$$

In this particular case, the estimator performs well, approaching the theoretical limit 0.421. This is despite the naivety of separately estimating four different entropy terms; dealing with this difficulty in a sophisticated way was the main breakthrough of the KSG estimator.

According to the stationary increments estimator, each estimated entropy term should only depend on the length T of the time window $[t_0, t_0 + T]$ we consider. In the following, the above are systematically utilized to derive pointwise estimates for cTE based on the data generated.

We are interested in estimating the cTE values shown in Figure 4.1. Thus, we begin by fixing the starting timepoint $t_0 = 0$ and consider the increasing time windows $[t_0, t_0 + T]$ for $T \in \{1, ..., N-1\}$, where N is the sample size of the data generated from the system. So, we consider the (discrete) increasing time windows $\{0, 1\}, \{0, 1, 2\}, ..., \{0, ..., N-1\}$. At every time window, we use all data available (e.g. $X_0, X_1, X_2, Y_0, Y_1, Y_2$ for $\{0, 1, 2\}$) to get a cTE estimate as described above. Thus, we map all time windows $\{0, ..., k\}$ for $k \in \{1, ..., T\}$ to a cTE estimate. This is the pointwise cTE estimate at time k, to be compared to the corresponding exact cTE value. For the specific model realization shown in Figure 4.6, Figure 4.7 contains the exact and the estimated cTE values overtime.

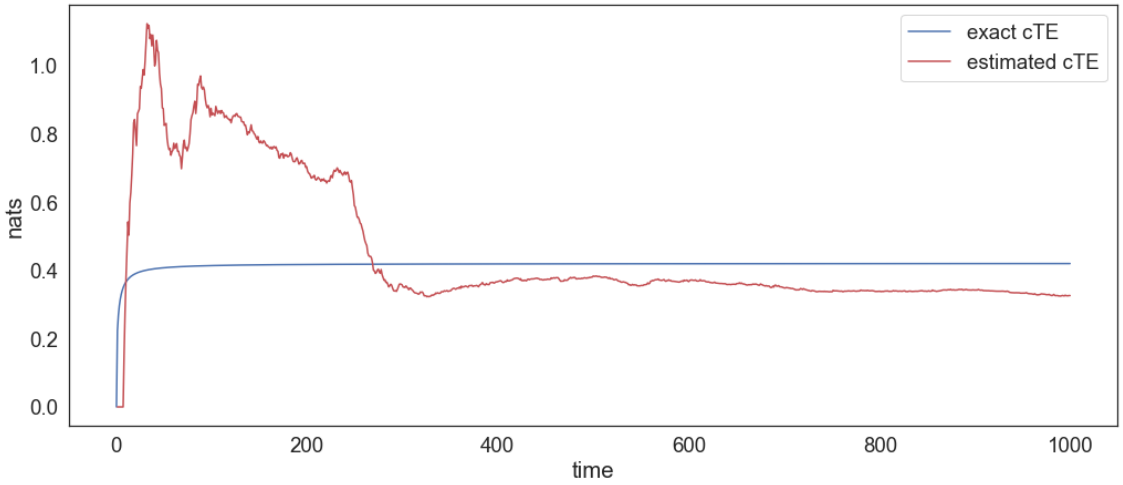

Figure 4.7: The exact cTE values and the estimated cTE plotted together.

The window size chosen for cTE estimation appears to be very important. While initially the estimator significantly overestimates the correct value, when a moderate amount of data have been made available, the estimates closely approach the corresponding exact value. It is important to note that once a moderate window size is reached the estimates stop improving, even showing a slow diverging behavior. This was also the case in other realizations, further illustrating the importance of a balanced window size for getting good estimates and the importance of correcting the estimates derived for bias.

Chapter 5

Data

This chapter presents the data that are used for the purposes of developing a benchmark framework for causal inference methods, referring to the second research question of the project. There are two different kinds of data included, simulated data and real data. In selecting any dataset to include in the project, knowing its causal structure is naturally very important - since only then we are able to evaluate the performance of any causal inference method. For simulated data, the corresponding causal structure can be inferred by the equations that generate the dataset.

5.1 Preliminaries

Benchmarking the performance of different causal inference methods in time series data is a task that was recently undertaken by many researchers, e.g. Runge et al. (2019a), Siggiridou et al. (2019), Papana et al. (2013), Krakovská et al. (2018), Kořenek and Hlinka (2018). Studying these papers, the researchers frequently utilize the following datasets to evaluate causal inference models on: vector autoregressive (VAR) systems, coupled logistic maps, Rössler systems, Lorenz systems, and Hénon Maps. Each dataset features its own theory, utility and causal structure.

The simulated dataset used for the analysis that follows is a multivariate adaptation of the Hénon Map. Historically, the introduction of the Lorenz system inspired the introduction of the Hénon Map so the former is discussed first.

Lorenz system

Among the systems mentioned above, the Lorenz system particularly stands out. It is a deterministic system consisting of three coupled ordinary differential equations. The following are based on Lorenz (1963) and Sparrow (1982).

Definition 5.1.1 (Lorenz system)**.** *The Lorentz system is defined as the following set of coupled ordinary differential equations:*

$$\frac{dx}{dt} = \sigma(y - x) \tag{5.1}$$

$$\frac{dy}{dt} = x(\rho - z) - y \tag{5.2}$$

$$\frac{dz}{dt} = xy - \beta z \tag{5.3}$$

with the initial conditions $(x(0), y(0), z(0)) = (x^*, y^*, z^*) \in \mathbb{R}^3$*, where* $\sigma, \rho, \beta, t > 0$*.*

Lorenz introduced the above system as a simplification of the convective motion of a fluid. As a mathematical object, it possesses several extraordinary properties. To motivate the discussion, Figure (5.1) features a solution of the system for specific parameter values.

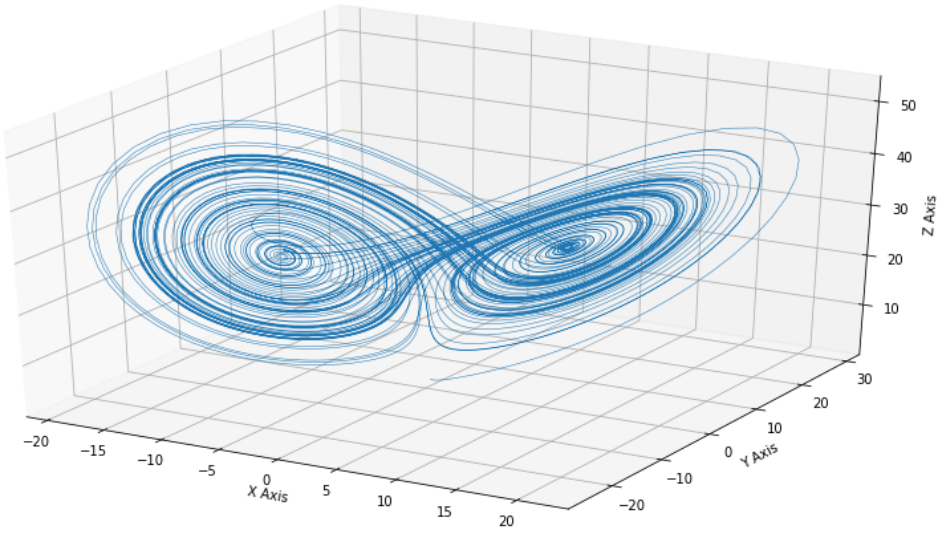

Figure 5.1: A numerically computed solution of the Lorenz system where $\sigma = 10, \rho = 28, \beta = 8/3$

For arbitrarily big t, the solution appears to be "trapped" in the general shape that is seen. Moreover, while the general form of the graph appears to be independent of the system's initial conditions, its numerical details are highly dependent on both the initial conditions of the system and the values of the parameters. In particular, the exact sequence of the values $(x(t), y(t), z(t))$ (i.e. how the solution traverses its graph) is extremely sensitive to changes in the initial conditions of the system. That is, no matter how close (in $\mathbb{R}^3$) two initial conditions are, the solutions of the same Lorenz system produced by them will ultimately have no connection, evolving independently from one another. This indicates that, in reality, making distant predictions of how such systems will develop will not be possible (Broer and Takens, 2011, p. 49).

Remark. *Lorenz, who was also a meteorologist, attributed the inaccuracy of long-term weather prediction to a similar sensitive dependence of the Earth's atmosphere on initial conditions. Illustrating this thesis, he used the example of the wing beat of a butterfly potentially "causing" a tornado as demonstration of a minuscule change in the initial conditions of a system leading to a significantly different future for it; a paradigm now known as the "butterfly effect" Lorenz (1995).*

Furthermore, recall that the system introduced in 5.1.1 is fully deterministic; there are no random variables involved, yet the system behaves in a "chaotic" and unpredictable way. Combining the above results and observations, we infer that it might be possible to model a highly complicated system through a simple deterministic system of low dimension - such as the one introduced by Lorenz. Achieving a highly complex structure with relatively simple means can be advantageous and therefore explains the popularity of Lorenz systems across many scientific fields, as well as the reason why researchers frequently choose to evaluate causality methods on them.

Causal graphs

Generally, by looking at the equations that define a multivariate coupled system such as the one in 5.1.1, we can infer a corresponding graph capturing the relations between variables. This graph contains one node for each variable of the system, and a directed edge from variable X to variable Y if and only if the equation where Y is defined depends on X. Thinking of the equation that defines a variable X as modelling the *causal mechanism* between this variable and the rest, we may call this graph *causal*, and interpret any system of such equations as a *causal model* carrying a specific *causal structure* Rubenstein et al. (2018).

For the example of the Lorenz system defined in 5.1.1, we observe the following relations between the variables:

- From 5.1, we see that (the derivative of) variable x depends on variables x and y
- From 5.2, we see that (the derivative of) variable y depends on variables x, y and z
- From 5.3, we see that (the derivative of) variable z depends on variables x, y and z

Note that in all three equations that define the (derivative of) each variable we also find the variables themselves. Thus, the variables of the Lorenz system are also self-dependent. From the perspective of causality this might be informally referred to as *self-causation*. The above remarks are now conveniently summarized in the Lorenz causal graph:

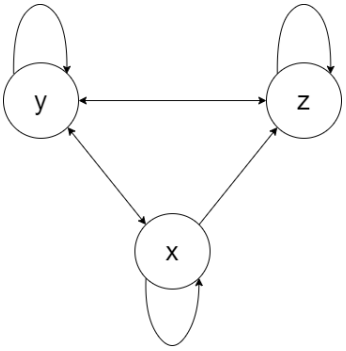

Figure 5.2: The Lorenz causal graph

Remark. *Note that, due to a directed edge from variable X to variable Y existing in the causal graph if and only if the equation where Y is defined depends on X, directed edges in a causal graph exclusively signify direct causation. In the above example, the path $Z \to Y \to X$ exists, and therefore Z is causing X, albeit in an indirect way (i.e. through Y). Therefore, the directed edge $Z \to X$ is not included in the graph.*

Hénon Map

In the Lorenz system, the dependence of each variable on the rest is formulated through the derivative of the variable, and not through the variable itself. This differential equation modelling might be fruitful in modelling physical phenomena, but it is not necessarily needed for the purposes of causal inference. A different mathematical object, named the *Hénon map* Hénon (1976), provides similar utility to the Lorenz system, without requiring to be defined in terms of derivatives.

Definition 5.1.2 (Hénon Map)**.** *The Hénon map is the following set of coupled equations that are defined recursively as:*

$$x_{t+1} = 1 - ax_t^2 + y_t \tag{5.4}$$

$$y_{t+1} = bx_t \tag{5.5}$$

where the system is initialized from $(x_0, y_0) = (x^*, y^*) \in \mathbb{R}^2$, the parameters $a, b \in \mathbb{R}$, and $t \in \mathbb{N}$.

The Hénon map was originally introduced with the aim of finding a system, which is as simple as possible while retaining similar properties to the Lorenz system - hence alleviating the large computational obstacles that a system of coupled differential equations posed at the time, and simplifying theoretical analysis.

The parameter values for a, b that are traditionally studied are $a = 1.4$ and $b = 0.3$. This parameter selection leads to the *classical Hénon map* - which indeed exhibits chaotic behavior similar to the Lorenz system Gonchenko et al. (2005).

Recall that solutions of Lorenz systems become "trapped" in a shape of the form (5.1). This also holds for the Hénon map, as its values quickly end up traversing a shape such as the one featured in (5.3). This graph possesses similar properties to the corresponding graph derived by solving the Lorenz system Gonchenko et al. (2013).

Specifically, the evolution of the Hénon map depicted in figure (5.3) was proven to never tend to a periodic pattern Benedicks and Carleson (1991), which has implications regarding its predictability (Broer and Takens, 2011, Chapter 2) - hence imitating a characterizing property of the Lorenz system. In the case of Lorenz systems, the convergence of a solution happens independently of the initial conditions. This is not the case for Hénon maps. In fact, depending on the initial values (x^*, y^*), the Hénon map will either end up traversing such a shape, or diverge to infinity. Potential diverging behavior can be easily discerned from (5.4), as initializing large values for x^* will lead to the quadratic term x_t^2 dominating the system, leading to divergence.

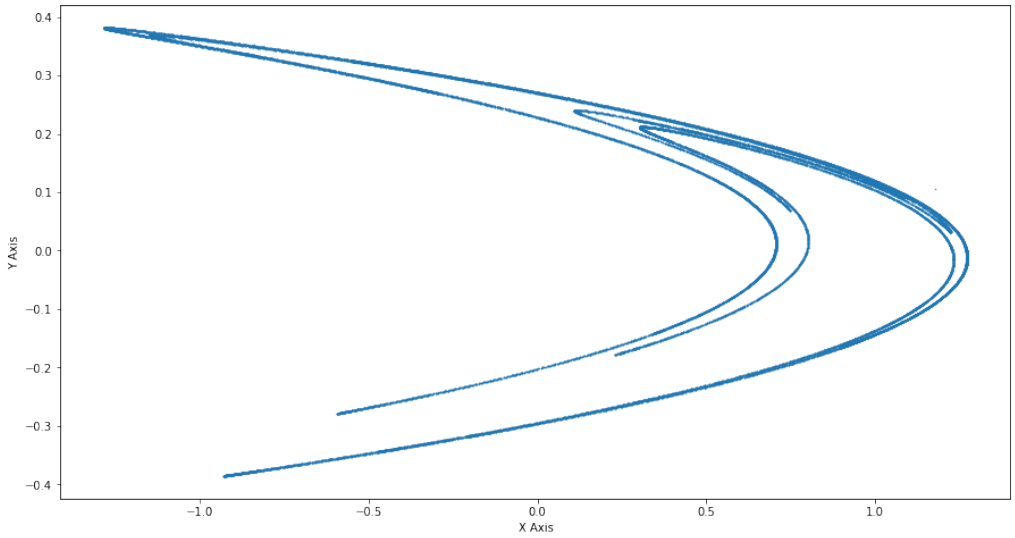

Figure 5.3: Plot of the first 40,000 iterations of the classical Hénon Map for $(x_0, y_0) = (0.35, 0.65) = \big((1-b)/2, (1+b)/2\big)$

Generalized Hénon map

For the purposes of the project, simulating a (mathematically simple) deterministic system that exhibits highly convoluted behavior is important, and the Hénon map is a great candidate. However, the original Hénon map formulation defined in (5.4) comprised a bivariate system. From a causal inference standpoint, a multivariate extension of the Hénon map is clearly needed, since in a bivariate setting only two possible causal relations between the variables exist.

This multivariate extension, called the *generalized Hénon map* was initially given in Baier and Klein (1990). This system consists of the following K variables, defined through the equations:

Definition 5.1.3 (Generalized Hénon Map)**.** *The generalized Hénon map is defined as the following set of K recursive equations:*

$$\begin{aligned} x_{i,t+1} &= a - x_{K-1,t}^2 - b x_{K,t} && \text{for } i = 1 \\ x_{i,t+1} &= x_{i-1,t} && \text{for } i = 2, ..., K \end{aligned}$$

where the system is initialized from vector $(x_1^*, x_2^*, ..., x_K^*) \in \mathbb{R}^K$. Here, $2 \leq K \in \mathbb{N}$, $a, b \in \mathbb{R}$, $0 < |b| < 1$ and $t \in \mathbb{N}$.

While important for our goals, this extension is rather limited. This can be seen by investigating the causal graph (Figure 5.4) of this system.

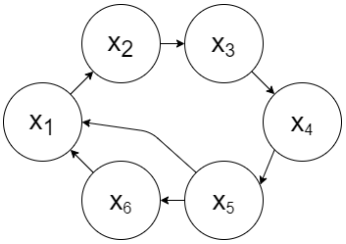

Figure 5.4: The generalized Hénon map causal graph for $K = 6$

Indeed, the equations defining the generalized Hénon map indicate a very simplistic causal structure (cyclical, one-directional, no self-causation) that is not satisfactory for accommodating the degree of complexity needed.

This is why the approach presented in Siggiridou et al. (2019) is followed. In this paper, a set of K modified Hénon equations is listed (originally studied in Schiff et al. (1996)), leading to a more involved causal structure to be picked up by any causal inference method. The definition of this system follows.

Definition 5.1.4 (Generalized Hénon maps for causal inference)**.** *The generalized Hénon map to be used for causal inference is defined via the following K equations:*

$$\begin{aligned} x_{i,t} &= 1.4 - x_{i,t-1}^2 + 0.3x_{i,t-2}, && \text{for } i = 1, K, \\ x_{i,t} &= 1.4 - \left(0.5C\left(x_{i-1,t-1} + x_{i+1,t-1}\right) + (1-C)x_{i,t-1}\right)^2 + 0.3x_{i,t-2}, && \text{for } i = 2, \ldots, K-1 \end{aligned}$$

where the system is initialized from vector $\left((x_{1,0}^*, x_{1,1}^*), (x_{2,0}^*, x_{2,1}^*), ..., (x_{K,0}^*, x_{K,1}^*)\right) \in \left(\mathbb{R}^2\right)^K$, $3 \leq K \in \mathbb{N}$, $t \in \mathbb{N}$ *and* $C \in (0, 1)$

Notice that the classical values $a = 1.4$ and $b = 0.3$ are used. In this definition, the parameter C denotes the *coupling strength*, i.e. how "strong" each causal relation is. This is conceptualized as a coefficient in front of the causal variables - with larger values of C indicating a tightly coupled system with a clearer causal structure.

Similarly to the original bivariate Hénon map, we choose to uniformly sample all initial values of the system from the interval $(0, 1)$. Following Kugiumtzis (2013), the coupling strength C is empirically restricted to the interval $(0, 0.8]$. We finally consider the corresponding causal graph of this system, for the case $K = 10$.

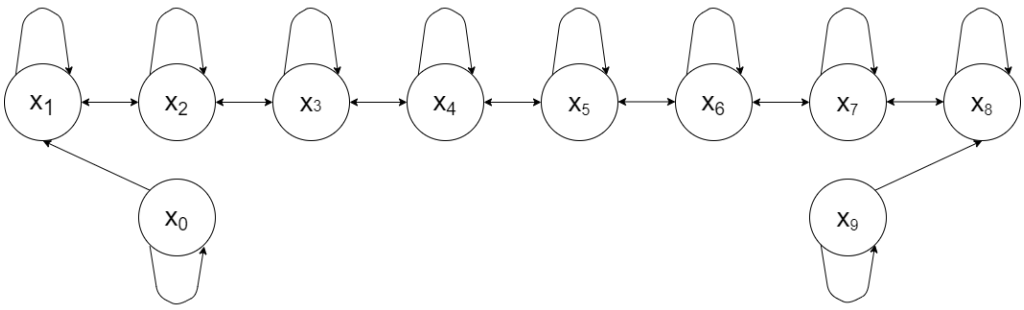

Figure 5.5: The modified generalized Hénon map causal graph for $K = 10$

5.2 The Hénon map

A dataset with $K = 10$ variables and $n = 1000$ observations each is generated from equations (5.1.4) at a uniform and constant sampling rate with coupling strength $C = 0.5$. The system is initialized from the point $\big((x^*_{0,0}, x^*_{0,1}), (x^*_{1,0}, x^*_{1,1}), ..., (x^*_{9,0}, x^*_{9,1})\big)$ that is uniformly sampled from $(0, 1)^{10}$. Each variable is a time series, and the time series are causally related as shown in Figure 5.5.

For the study that follows, this dataset will be re-generated with different variable numbers, coupling strengths and initial conditions. Similarly to the Lorenz system, minute changes in the initial conditions of the system will lead to significantly different time series but the system will retain its macroscopic features. Varying the coupling strength C will affect the level of influence that variables have on each other and will therefore alter the "difficulty" causal inference methods will deal with. The aforementioned instance of the dataset is chosen here only for visualization and exploration purposes. The Hénon Map dataset is visualized in Figure 5.6.

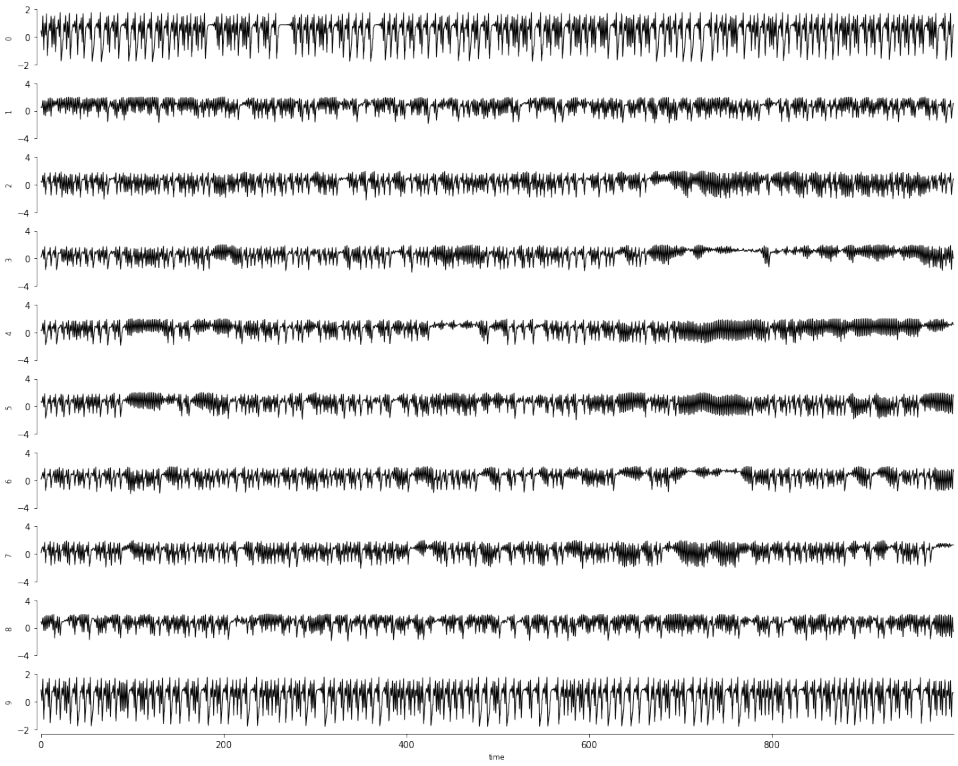

Figure 5.6: 10 time series of the Hénon map plotted over the same axis

The dataset appears to exhibit intense oscillations. Hypothesizing about any causal relation between its variables by simply examining the above plots is not possible. To further explore the dataset, a correlation matrix of the variables is shown in Figure (5.7).

Since the main diagonal consists of correlations of each variable with itself, it is equal to 1. The general pattern observed in the correlation matrix is the existence of positive correlations between successive variables. A part of the causal truth is therefore transferred here, which is mostly attributed to the multitude of linear relations existing between the variables in (5.1.4). Nevertheless, due to the correlation coefficient being a symmetric measure, a notion of directionality between two variables can not be derived.

Moreover, correlation strength of successive variables varies. Variables 3 and 4 correlate moderately, while a strong correlation is detected between variables 7 and 8, potentially confounding inferences. Notice that positive correlations also exist between non-successive variables, e.g.

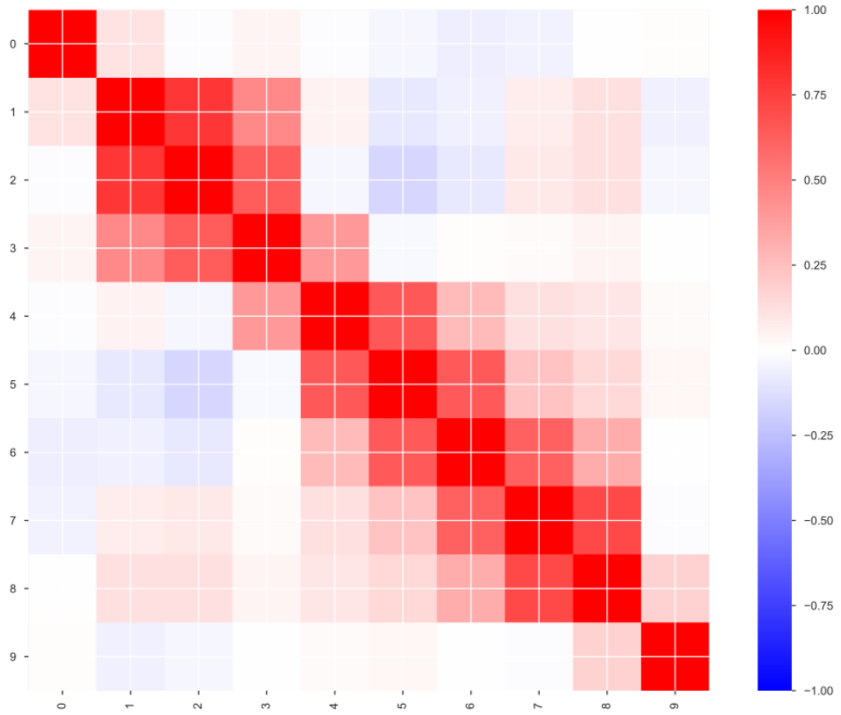

Figure 5.7: Pearson's ρ correlation coefficient for every pair of variables in the Hénon Map dataset.

between variable 3 and 5. This is due to *indirect causation*, illustrated by the dependence of the equation defining variable 5 on variable 4, which in turn depends on variable 3, i.e. by the existence of the path $3 \to 4 \to 5$ in the causal graph of the system. Many causal inference methods account for indirect causation, in which case the causal relation $3 \to 5$ should be deemed insignificant.

5.3 Real data

The following ASML dataset will also be used for benchmarking the performance of causal inference methods. It consists of three univariate time series named $P1, P2, P3$ shown in Figure 5.8.

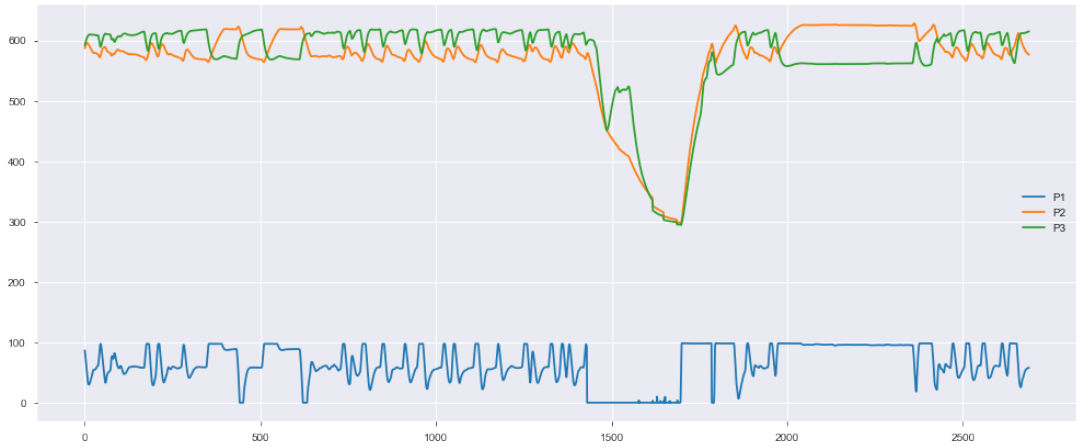

Figure 5.8: Real data consisting of three time series

All three time series consist of 2689 datapoints sampled at a uniform and constant frequency. The causal relations between these three time series are known by design. It is important to note that there exist unobserved variables influencing this dataset, something that will generally compromise causal inferences. This is normally the case in real data, therefore benchmarking causal inference methods in such a dataset is realistic. A causal inference method might be able to remedy the effect or hypothesize the existence of such exogenous variables. The causal structure of the system is visualized in the corresponding causal graph included in Figure 5.9.

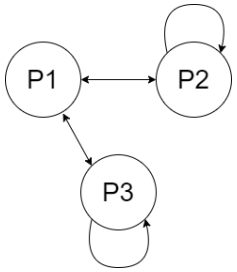

Figure 5.9: Real data causal structure. In reality, this is a subgraph of the full causal structure graph, as the existence of at least one time series influencing both $P2$ and $P3$ was confirmed by domain experts. However, this is not observed.

Moreover, all three time series exhibit stationary behavior, with the exception of an interval where a large drop in the values of $P2$ and $P3$ occurs. In Chapter 6 we will see that many causal inference methods assume stationary data; for the purposes of the benchmark study, we will consequently focus on the stationary time interval $[650, 1400]$. Stationarity of the time series within this interval is also substantiated by the execution of an augmented Dickey-Fuller test Said and Dickey (1984) that rejects the presence of a unit root for all time series.

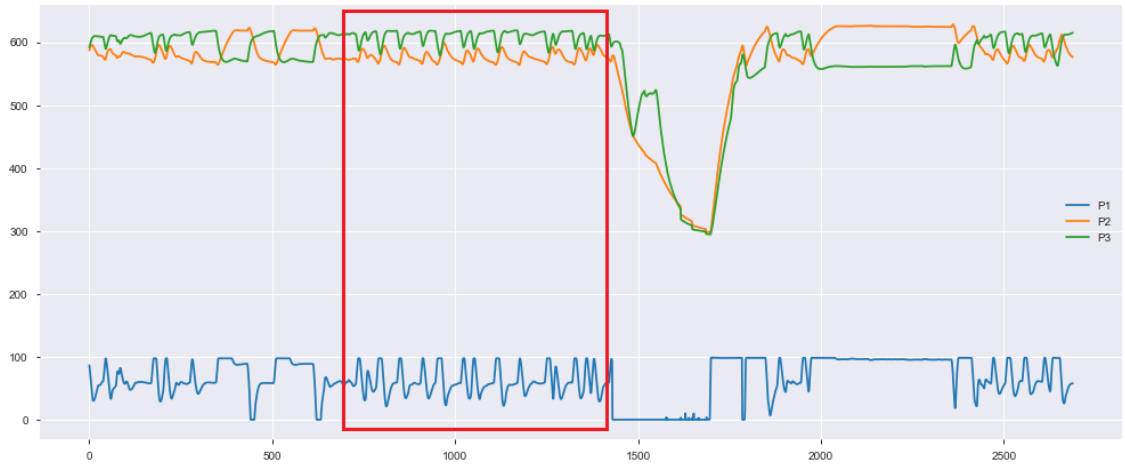

Figure 5.10: The data window to be used in the study is highlighted in red

Chapter 6

Benchmark Framework

This chapter introduces a benchmark framework for the study, evaluation, and comparison between different causal inference methods in time series analysis. The framework consists of three main components: data used (also discussed in Chapter 5), methods selected, and finally, the performance evaluation and comparison between these methods. In this chapter, the overall structure and methodological details of the framework are presented, alongside a comprehensive discussion of each method used. Details regarding the performance evaluation and comparison are also part of this chapter; however the full results, the insights derived from them, and their interpretation are included in Chapter 7.

As it was also noted in Chapter 5, several similar frameworks have been recently developed (e.g. Runge et al. (2019a), Siggiridou et al. (2019), Papana et al. (2013), Krakovská et al. (2018), Kořenek and Hlinka (2018), Nauta et al. (2019)). Each study investigates a variable number of methods over multiple datasets with known causal structure via performance measures. For the development of the current framework, the overall methodology used in each of these papers was consulted. While simultaneously minding the particular interests of ASML, this led to the framework structure proposed and outlined in this chapter. Overall, the approach developed for this project is mostly influenced by Siggiridou et al. (2019) and Nauta et al. (2019).

6.1 Goal

As it was mentioned before, in this project, the focus is on unveiling the causal structure of a dataset (in contrast to deriving insights related to causal effects, or counterfactual questions). Here, we elaborate on the meaning of this task. So, the goal of the benchmark study is set.

A general question, not only of causal inference, but of multivariate data analysis in general, is the following: Given a dataset with M variables, is it possible to infer a graph with M nodes, where edges indicate that two variables are "related"?

From the perspective of the project, variables are time-dynamic and "relations" are causal. A specific notion of causality according to Granger was described in Chapter 2. While this is the most popular causality notion for time series, we note that not all methods included in the study are precisely based in it (e.g. recall the subtle details presented for TE on Section 2.2.2). When causality is assigned a significantly different meaning within a method, we shortly elaborate on it in the same section where the method is presented.

Inspired by terminology used in neurosciences (Friston (2011), Sporns (2010)), Bossomaier et al. (2016) use the term *effective network inference* to describe the task of retrieving a directed graph encoding time-lagged causal interactions between the variables from a given dataset. This is contrasted to structural or functional network inference, the former relying on interventional techniques to infer *physical* causality Pearl (2000), Ay and Polani (2008) and the latter being based on correlation analyses.

From the discussion of what causality generally refers to in this project, it is evident that

the goal of this benchmark study is to perform *effective network inference* in data consisting of interacting time-series. All causal inference methods presented here attempt to deal with this task. This is visually explained in Figure 6.1.

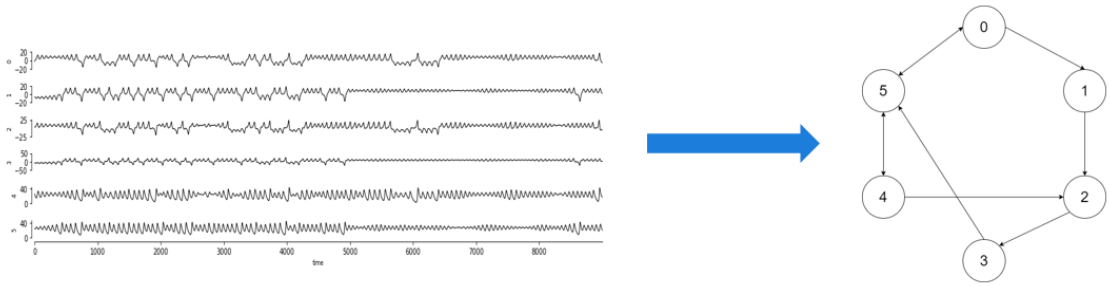

Figure 6.1: Effective network inference: in the directed graph, each directed edge denotes a time-lagged causal interaction.

As a familiar example, transfer entropy has been successfully used for effective network inference in a variety of fields, and is moreover recognised by researchers as a natural option for such a task. This is related to its theoretical background, as shown in Chapter 2: recall that TE is measuring a directed relationship between variables in terms of the uncertainty of a target that is resolved by a source, a notion that is associated with Granger causality.

However, TE is not the only method designated for this task. In fact, within Information Theory only, there exist similar notions aiming at solving the same problem Massey (1990), Vlachos and Kugiumtzis (2010). Of course, the search for other methods need not be constrained to Information Theory, and multiple methods from diverse theoretical backgrounds exist. Selecting, evaluating, and comparing the performance of several methods in a homogenized and objective manner, while minding the methodological details is therefore the goal of this benchmark study.

6.2 Data

So far, preliminary knowledge was provided and exploratory data analysis was performed in data to be used in Chapter 5. This section features specific details of the data used from the perspective of this benchmark study. We first summarize the properties of data used in the following table.

	Hénon Map	Real Data
Data	continuous	continuous
Number of variables	arbitrary	3
Data points	1000	2689
Time index	discrete	discrete
Stationarity	yes	partly

Table 6.1: Summary of data properties

6.2.1 Hénon map

In Chapter 5, we simulated a generalized Hénon map dataset with a causal structure as shown in Figure 5.5. A change in the initial conditions that span the dataset results in significantly different time series, allowing us to generate different datasets by simply randomizing the initial conditions. Nevertheless, the causal structure of the dataset will remain the same, thus enabling the possibility of objective comparison of causal inference methods over these datasets.

The goal here is to iterate all causal inference methods presented below over an arbitrary number of different Hénon datasets. Recall that each Hénon dataset has three free parameters:

- The coupling strength C
- The number of variables M
- The number of datapoints each variable has N

To introduce proper variety, four different Hénon map data categories are introduced, varying the first two parameters listed above:

- H_1: A set of high dimensional datasets ($M = 20$) of high coupling strength ($C = 0.5$)
- H_2: A set of low dimensional datasets ($M = 5$) of high coupling strength ($C = 0.5$)
- H_3: A set of high dimensional datasets ($M = 20$) of low coupling strength ($C = 0.25$)
- H_4: A set of low dimensional datasets ($M = 5$) of low coupling strength ($C = 0.25$)

For the low dimensional categories H_2 and H_4, 20 different Hénon datasets are generated by randomizing the initial conditions of the system. Each dataset contains $N = 1000$ points, sampled at a constant and uniform sampling rate. Methods are subsequently iterated over all these $20 + 20$ datasets. For the high dimensional categories H_1 and H_3, due to restrictions imposed by computational resources, 5 different Hénon datasets are generated as before only. Methods are subsequently iterated over all these $5 + 5$ datasets.

The performance of each method at every iteration is then evaluated using the metric presented in Section 6.4 and its average value is calculated (both per category and overall by considering the arithmetic mean of the category averages). In parallel with benchmarking performance, the running time of each iteration is recorded, and its median value per category is reported. Averaging the median runtime over all categories, a single runtime per method is also reported.

6.2.2 Real data

As it was already mentioned in Chapter 5 only a specific time window of the real dataset is going to be used. This follows not only after requiring stationarity of data, but after the advice of a domain expert to also avoid 0 values in variable $P1$ as they are abnormal. As a result, 750 datapoints are left in the dataset and we therefore choose to run each method over all 750 datapoints only once. Equivalently to simulated data, the performance of each method is then evaluated and its running time is registered.

6.3 Methods outline

A wide variety of causal inference methods in time series exists in literature. For this project, two criteria regarding the inclusion of a method in the benchmark are deemed important:

- Method diversity
- Relevance to the context of ASML

With regards to the first criterion, we note that a similar framework may be developed consisting of techniques coming from e.g. Information Theory only, or techniques that are solely based on different implementations of Granger Causality. For this project, this is not the intended level of granularity; causal inference in time series does not have a long history either as a scientific discipline or in the industry. Therefore, the search for methods should remain broad enough, and no such constraints should be imposed on it.

Secondly, the company-specific context matters. Since every method to be considered comes from a specific theory, the advantages and limitations of the underlying theory are conveyed to the methods themselves. For example, methods that are based on Information Theory offer high flexibility as they do not assume a model for describing the data, but their estimation is generally challenging and computationally expensive. Thus, slow yet general methods and fast yet limited methods should both be studied. As another example, at ASML, a huge number of time series are registered over extraordinarily different time scales, ranging from nanoseconds to days. Instantaneous causation might considerably confound causal inferences, especially in the case of methods based on Granger Causality: interactions might happen so fast that the cause appears to no longer precede the effect in data, and the first axiom Granger assumed is violated. Thus, methods that are able to deal with instantaneous causality should also be included in the study.

Taking into account the above, the following 7 methods were selected for inclusion in the study. The following table summarizes them, and it also lists the theoretical background they belong to as well as the main reference for each with regards to the specific implementation used.

Abbr.	Method	Background	Ref.
MTE	Transfer entropy	Information Theory	Wollstadt et al. (2019)
PMIME	Partial MI on mixed embeddings	Information Theory	Kugiumtzis (2013)
MVGC	Multivariate Granger causality	Granger Causality	Barnett and Seth (2014)
PDC	Partial directed coherence	Granger Causality	Baccala et al. (2007)
PCMCI	PC algorithm with momentary CI	PC algorithm	Runge et al. (2019b)
CCM	Convergent cross mapping	Dynamical Systems	Clark et al. (2015)
TCDF	Attention-based CNNs	Neural Networks	Nauta et al. (2019)

Each method is later going to be presented separately in the current chapter. Details on how they were applied within the study and important properties will also be discussed.

6.4 Evaluation

Evaluating a causal inference method will happen over two axes. First, a list of relevant qualitative properties for causal inference methods is given below. During the presentation of each method, each property is discussed. The results are then summarized in Table 6.2. This section features a discussion of each property; they are all pertaining to causal inference questions, caveats and important things to consider while inferring causality.

The second axis for the evaluation of a causal inference method is related to measuring its quantitative performance. This is done via evaluation metrics that are also used in binary classification. This connection is also elaborated in the current section.

6.4.1 Qualitative properties and classification

Depending on the domain of application the selection of a method for inferring causality from a temporal dataset might range from incompatible to a suitable fit. Bielczyk et al. (2017) compile a similar list of important properties for causality methods from the perspective of brain sciences. For the context of this project, the following properties are important; they can also be used to characterize and classify the methods presented later.

Delay Discovery

As it can be noted from e.g. the equations that span the Hénon dataset, a *delay* between a cause and its effect may exist in a dataset. It is important to know, whether a method can retrieve this and relevant details.

Self-causation

A time series might be, at least partially, causing itself. For an example in the context of Granger causality, if a time series is useful in predicting itself as indicated by a good model fit of e.g. an autoregressive model then this time series can be thought of as "causing" itself. Self-causation is denoted by an edge starting from a node and ending on the same node in the causal graph a method retrieves. This might not be of central importance in inferring the causal structure of a dataset however some methods are able to detect it.

Instantaneous causality

Instanteneous causality exists in a dataset if the cause and the effect are reported at the same time, i.e. if the causal delay is 0. Ideally, the cause preceding the effect is a sound assumption to make. In practice however, due to computational limitations and erroneous sampling frequencies, instantaneous causality may arise.

(Unobserved) confounders and indirect causation

A formal definition of a confounder within structural causal models is given in Pearl (2000). Here, we call variable X a confounder of variables Y and Z whenever X is causing both Y and Z. This results in spurious correlations arising between Y and Z and is amongst the fundamental challenges in causality. In the figure below, X is a confounder for Y and Z.

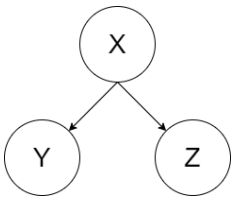

Figure 6.2: X is a confounder for Y and Z

Distinguishing direct from indirect causal effects is also of fundamental importance. A method should only report direct causal relations. In the figure below, X is causing Z indirectly, so the edge $X \to Z$ should not be part of the directed graph a method retrieves.

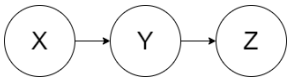

Figure 6.3: X is indirectly causing Z. The relation $X \to Z$ should not be detected.

An especially challenging case is when a confounder is not included in the dataset. This significantly increases the difficulty of excluding spurious associations from causal analysis. A method might be able to (at least) hypothesize the existence of unobserved confounders.

Polyadic relations

When using a directed graph where nodes are variables of a dataset and edges are interactions between the variables to model the structure of a system, a subtle assumption is made. Interactions between variables are assumed to be dyadic: that is, when the relation $X \to Y$ is inferred, causation of Y is uniquely the result of X, and not the result of a potential *synergy* between X and other variables that leads them to only *jointly* cause Z, i.e. the result of a *polyadic* relation. Such higher-order dependencies cannot be represented in a graph, unless additional nodes are used. Whether a method infers dyadic or polyadic relations is therefore an important property.

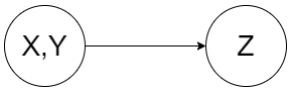

Figure 6.4: Causation of Z may be the result of a synergy (polyadic relation) between X and Y. X and Y considered separately might not be causing Z.

Non-linear relations

Non-linear patterns are frequently encountered in data as well as in the way variables interact. Depending on the assumption a method makes, it might not be able to capture non-linear relations within a dataset. This should be considered when selecting a method.

Computational complexity and network size

From a practical standpoint, the computational complexity of any method is key. The effect that an increase in the number of variables of a dataset has in the running time of a method is also very important. This might be related to subtle details of how a method estimates or calculates required quantities that might be adversely affected by high-dimensionality.

If the theoretical complexity of a method is available, it will be mentioned in the section where the method is presented. In any case, running times will be reported in Chapter 7 and computational complexity will be discussed there.

Bivariate / Multivariate data

A method might be suitable for application over an arbitrary number of variables simultaneously, or it might be designated for bivariate inferences each time. Bivariate methods may suffer from issues caused by confounders, but they are generally faster.

Discrete / Continuous data

Discrete data are generally more convenient to work with, in terms of e.g. estimation or speed. Whether a method is designed for discrete or continuous data should be acknowledged.

Stationarity

As it has been demonstrated so far in the report, time series analysis methods are very frequently assuming stationary data. Depending on the application context, a method internally accounting for potential non-stationary patterns might be considerably advantageous.

In addition to discussing the above properties, the following table reports whether certain properties hold for the data we consider.

	Hénon Map	Real Data
Self-causation	yes	yes
Confounders	observed	observed & hidden
Type of relations	non-linear	unknown
Causal delays	1-2	unknown

6.4.2 Quantitative performance evaluation

Referring to Figure 6.1, we note that at a mathematical level, given a temporal dataset, a causal inference method returns a directed graph. The important information this graph should convey is which directed edges exist. So, assume that we have M variables (time series) $X_1, ..., X_M$.

A causal inference method associates a binary value (existence/absence) to all directed pairs of variables we can create (there are M^2 such pairs). The evaluation of methods will happen by simply investigating the directed graph skeleton.

This remark, shows the direction for evaluating such causal inference methods: we should evaluate how "well" a method maps all possible directed variable pairs to 0 or 1. The advantage that a benchmark study carries, is the fact that we are aware of the causal mechanism that generated the data we gave to a method. That is, the causal *ground truth* directed graph that spanned the data is available (see Section 5.1 for details on how this graph is obtained).

To summarize, we therefore expand the context visualized in Figure 6.1 with another directed graph, constituting the ground truth that was used to generate the data.

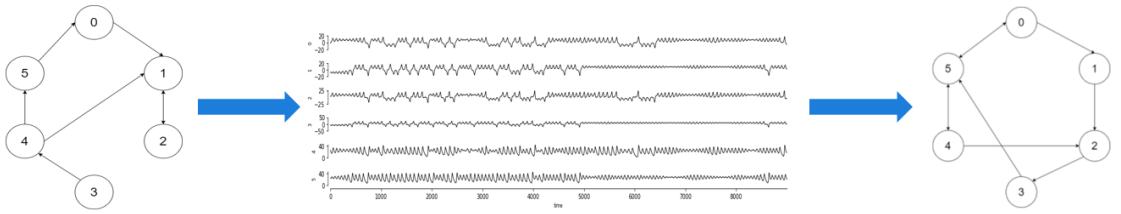

Figure 6.5: Data are generated from a system with known causal structure. Then they are provided to a causal inference method. Ideally, the method would return the initial directed graph.

Subsequently, note that any unweighted directed graph can be fully represented by its (binary) adjacency matrix. Thus, once we obtain the directed graph a causal inference method estimates, we may simply compare the adjacency matrix of this graph, with the corresponding ground truth adjacency matrix. Concatenating each row of both matrices, we obtain two binary vectors to be compared.

What was essentially described above, is the treatment of the evaluation of a causal inference method as the evaluation of a *binary classifier* over the set of all directed variable pairs of a dataset. The relevant literature of binary classification evaluation metrics can then be consulted to select the quantitative metric desired. A short description of such metrics is given here, followed by a discussion regarding metric selection.

Confusion Matrix

The confusion matrix contains the four fundamental quantities needed for binary classification.

- *True Positives*: number of 1's (existent edges) classified as 1 (existent)
- *False Positives*: number of 0's (absent edges) classified as 1 (existent)
- *True Negatives*: number of 0's (absent edges) classified as 0 (absent)
- *False Negative*: number of 1's (existent edges) classified as 0 (absent)

So, the confusion matrix is the following 2×2 matrix:

$$\textit{confusion matrix} = \begin{pmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{TN} & \mathrm{FN} \end{pmatrix} \tag{6.1}$$

Sensitivity

Sensitivity, otherwise known as recall or true positive rate measures the proportion of actually existent edges correctly identified as such:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6.2}$$

Specificity

Specificity, also known as true negative rate, measures the proportion of actually absent edges that are correctly identified as such:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6.3}$$

F1 score

The F1 score is given by:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6.4}$$

Matthews correlation coefficient

Finally, the correlation coefficient of Matthews (MCC) is the following:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6.5}$$

Discussion

While all metrics discussed above are informative in their own way, in order to later rank the performance of methods a specific metric should be used. Reviewing the literature of binary classification metrics as well as minding the specific context of the project, the correlation coefficient of Matthews is chosen as the main evaluation metric to be reported for each method.

The MCC was first introduced in Matthews (1975). A main advantage of MCC is the fact that it includes all 4 elements of the confusion matrix in its calculation. This is to be contrasted to the F1 score, which does not account for the performance of the method with respect to true negatives. MCC is easily understood, since its values vary between -1 and 1 (comparably to other correlation coefficients) with larger values indicating better performance. It was designed as a correlation measure between the actual and the predicted values the method yields.

MCC was evaluated as a performance metric in Powers (2011) and it is generally regarded as one of the best measures to summarize the confusion matrix with a single number. It favorably compares to the F1 score Chicco and Jurman (2020) also being appropriate for imbalanced data Boughorbel et al. (2017). The F1 score is only going to be reported for the real dataset, as disregarding true negatives allows for better comparisons in that particular case.

6.5 Methods

All methods included in the benchmark study are presented in this section. The focus is on concisely summarizing each method and commenting on its implementation details. This entails the experimental design as well as choices on the parameters chosen for each method. We also consider all properties that were included in Section 6.4 and discuss whether they apply to each method. This culminates in Table 6.2 where these properties are summarized for all methods, consequently proposing a classification scheme for causal inference methods in time series. We begin with (multivariate) transfer entropy.

6.5.1 MTE

As we saw in Chapter 3, the estimation of transfer entropy can be challenging. When this is combined with large and potentially high-dimensional datasets, where interactions between variables occur and should be therefore accounted for when inferring causality, we may expect significant computational challenges in employing transfer entropy for the analysis of a dataset.

Thankfully, since TE has found successful applications in a wide variety of practical settings, there have been notable advances through specific algorithms that aim to reduce the computational load. The algorithm that is used to estimate TE and analyze a dataset with it, is based on Faes et al. (2011) and Lizier and Rubinov (2012). Wollstadt et al. (2019) implemented this algorithm to perform effective network inference with TE.

Method Presentation

Assume a dataset consisting of M time series/variables $X = \left(X_t^{(1)}, ..., X_t^{(M)}\right)$ (in discrete time). In order to infer a directed network as described in Section 6.1 with TE, a brute force approach would serially consider each variable of the dataset $X^{(j)}$ as the target, and then estimate the transfer entropy from *all* potential sources $X^{(i)}$ of the dataset, one at the time, through appropriately selected embedding vectors for the target and each source.

Subsequently, we could test the significance of the estimates and retrieve all significant causal relations $X^{(i)} \to X^{(j)}$. However, in a multivariate setting interactions between time series might exist. Therefore, when examining any transfer entropy, such interactions must be taken into consideration by *conditioning* them out. So, for each target $X^{(j)}$ and source $X^{(i)}$ we would compute the transfer entropy of the following conditional relation:

$$(X^{(i)} \to X^{(j)})|(X \setminus \{X^{(i)}, X^{(j)}\}) \tag{6.6}$$

We would likewise estimate these transfer entropies and assess their significance by utilizing methods described in Chapter 3. Then, in theory, the full directed network is retrieved.

In practice, it is clear that such a brute force approach would be intractable. Besides considering *all* possible directed pairs $(X^{(i)}, X^{(j)})$, the addition of *all* other variables on the conditional part of transfer entropy creates high-dimensional quantities that are very challenging to estimate.

The main idea of the algorithm used is to restrict the conditioning set by finding a set of "relevant" sources for each target. The algorithm iteratively builds this set (starting from the null set) through a greedy approach, while extensively testing for significance.

Experimental Design

The multivariate transfer entropy algorithm presented in Wollstadt et al. (2019) was used. The required (conditional) mutual information quantities were estimated using the KSG estimator. The "relevant" source variables for each target were selected from the last 3 values of each source.

Qualitative properties

The algorithm used is able to discover causal delays, doing so by locating the relevant variable that maximizes a mutual information statistic. It does not report self-causal relations, while it can be adjusted to infer instantaneous causations. As a multivariate method, it is able to recognize (non-hidden) confounders, and distinguish direct from indirect relations. As a model-free information theory based method, it is very robust to non-linear patterns in the data. It is of course able to accommodate an arbitrary number of discrete or continuous variables, and a bivariate (unconditional) TE implementation is also offered within the same implementation. Stationarity is assumed, since it is required by the KSG estimator used. TE is only able to detect dyadic relations (see Section 2.2.2), however a partial information decomposition analysis Williams and Beer (2010) is provided.

For this particular code implementation, the authors provide the theoretical complexity of the algorithm, stating that the number of calculations required scales with $\mathcal{O}(M^2 \cdot d \cdot \ell \cdot S)$, where M is the number of variables, d is the average in-degree of the inferred network, ℓ is the maximum (user-defined) temporal depth search for each source, and S is the number of the surrogate calculations that are used for significance testing. We observe that the dimensionality of the dataset is the most significant factor in the speed of the method, a frequent problem in information theory and TE in particular Runge et al. (2012).

6.5.2 PMIME

The second information theory method considered is named PMIME, an acronym for Partial Mutual Information from Mixed Embedding. It was proposed in Kugiumtzis (2013) as a measure to perform effective network inference in multivariate time series. PMIME builds upon the MIME method proposed in Vlachos and Kugiumtzis (2010) for bivariate analyses, extending it to the multivariate case.

Method Presentation

Similarly to transfer entropy, PMIME is a method based on conditional mutual information. It was designed to alleviate several problems plaguing transfer entropy, such as issues stemming from potentially high-dimensional data, the need to wisely select the embedding parameters as well as intricacies related to significance testing of estimates. By circumventing these problems, PMIME aims to have significantly smaller computational complexity and a more straightforward implementation.

Consider M time series $\{X_t, Y_t, Z_t^{(1)}, Z_t^{(2)}, ..., Z_t^{(M-2)}\}$. The notation was chosen as such to easily denote the conditional relation we are interested in:

$$X \to Y|Z \tag{6.7}$$

where $Z = (Z^{(1)}, Z^{(2)}, ..., Z^{(M-2)})$. Vlachos and Kugiumtzis (2010) introduced a non-uniform embedding scheme for time series that bypasses the selection of embedding parameters via the use of information criteria regarding the past, current and future states of the time series. This allowed the authors to derive a measure for the strength of a bivariate directional coupling $X \to Y$ (having a similar meaning to $TE_{X\to Y}$), termed Mutual Information from Mixed Embedding (MIME).

Kugiumtzis (2013) subsequently extended this measure to the aforementioned multivariate case by first applying MIME on the joint space of X, Y, Z hence deriving a non-uniform embedding vector W of the joint process that "best explains the evolution of the target Y". This vector only contains the most "relevant" past states of all variables included, minimizing the dimension of the problem. It is obtained through iterative augmentation, starting from the null set and locating past variables that maximize the information gain for the future of the target, stopping when a specified threshold is reached.

The presence of past states of X in this vector signifies that X influences the evolution of Y, in which case, by construction, the corresponding PMIME measure is positive. Absence of past states of X from this vector indicates the lack of an influence, and PMIME is exactly 0. Because of this property, significance testing is not required. PMIME is defined as the following quantity:

$$R_{X\to Y|Z} = \frac{I(Y_t^T; W_t^X | W^Y, W^Z)}{I(Y_t^T; W_t)} \tag{6.8}$$

Here, Y_t^T is a vector containing T future values of the target, $Y_t^T := (Y_{t+1}, ..., Y_{t+T})$. Also, W_t is the non-uniform (mixed) embedding of the joint variable space containing the most relevant past states of all variables involved, with the sets of variables X, Y and Z that belong to W_t denoted by W_t^X, W_t^Y and W_t^Z respectively.

Experimental Design

The number of nearest neighbors for CMI estimation was set to 4. The past window size to search for significant CMI's was 5, while the future values array consisted of the immediate future only ($T = 1$). Following the advice of the authors a fixed threshold of 0.03 for PMIME termination was used.

Qualitative properties

As a multivariate algorithm based on mutual information, PMIME shares multiple properties with multivariate TE: it can deal with non-linear data and it is able to discern (non-hidden) confounders and distinguish direct from indirect relations. As implemented by the authors, PMIME focuses on detection of causal relations, and not on calculation of causal delays. It will also not report self-causal nor instantaneous or polyadic relations. Stationarity in data is assumed, and it can be applied to an arbitrary number of discrete or continuous time series.

6.5.3 PCMCI

Earlier in the report, we saw how TE encodes a specific conditional independence relation between time series. Conditional independence is in general very important in inferring causality. In fact, a wide class of conditional independence (CI) algorithms are popular in causal inference Spirtes et al. (2000). PCMCI (PC algorithm with Momentary Conditional Independence) is a recently developed algorithm for causal inference in time series; it is based on the popular PC algorithm broadly used in causal inference. The PC algorithm is a versatile method that can use several tests for CI between variables, both parametric and non-parametric, each with different advantages and disadvantages. The algorithm begins with a skeleton phase, where causal adjacencies between variables are detected through testing for CI, followed by orientation phases where the causal links become directed Runge (2020).

Method Presentation

The PCMCI algorithm Runge (2018) begins by applying the PC algorithm in order to identify a set of relevant past states for each time series and then uses a *momentary conditional independence* (MCI) test to evaluate whether causal relations exist within this set and the present state of all time series.

In short, let $X_t = (X_t^{(1)}, ..., X_t^{(N)})$ be N time series. The first step of the algorithm considers the present state of each time series $X_t^{(j)}$ and an embedding set of the N-dimensional vector X_t that contains the *preliminary parents* (sources) of $X_t^{(j)}$:

$$\widehat{\mathcal{P}}(X_t^{(j)}) := (X_{t-1}, ..., X_{t-d}) \tag{6.9}$$

This set of preliminary parents is iteratively thinned via a series of conditional independence tests, and the algorithm converges if no more tests can be executed. By that point the set $\widehat{\mathcal{P}}$ only includes the true causal parents of $X_t^{(j)}$ potentially alongside some false positives.

During the second step of the algorithm, the remaining variables of set $\widehat{\mathcal{P}}$ are evaluated with a MCI test and conditional independence is established. Both stages of the algorithm can be paired with linear or non-linear and non-parametric tests.

Experimental Design

The "ParCorr" conditional independence test implementing linear partial correlation was used. The null distribution of the test statistic was assumed to be Student's t. The embedding dimension for the set of preliminary parents was set to $d = 3$. The significance level for the first step of the algorithm was left unspecified, and it was internally optimized for each time series by the algorithm.

Qualitative Properties

PCMCI is able to discover causal delays, potentially reporting multiple delays for each causal relation. It also infers self-causal relations. The current software implementation does not accommodate instantaneous causalities, however a recent extension aims to fill this gap Runge (2020). The absence of unobserved confounders is among the main assumptions of the method. Provided confounders are observed, the algorithm is able to deal with them, and it also accounts for indirect

effects. All relations detected are dyadic. Depending on the type of CI test used, the algorithm can also deal with non-linear relations within the dataset. We deliberately chose a linear CI test, as the non-parametric option (based on mutual information) is significantly slow. The main advantage of the linear CI test selected is its computational complexity, which is orders of magnitude smaller compared to alternative multivariate methods. The method can deal with discrete and continuous data, and stationarity is assumed.

6.5.4 CCM

Convergent Cross Mapping was introduced in Sugihara et al. (2012). The time series X_t and Y_t are assumed to belong to two underlying dynamical systems X and Y, respectively. Due to Takens (1981), we are able to reconstruct (up to a diffeomorphism) the state-space of the dynamical systems by two manifolds M_X, M_Y given by the delay embeddings $X_t^{(d,\tau)}, Y_t^{(d,\tau)}$. This implies that the manifolds share certain properties with the corresponding state space. Notably, points that are close in the original state-space are also close in the reconstructed manifold, and the neighborhoods of any point are preserved. Causality is then detected from system X to Y (i.e. from time series X_t to time series Y_t) if the time indices of nearby points in the past data of M_Y are able to identify nearby points in M_X Krakovská et al. (2018).

Method Presentation

Specifically, the CCM algorithm works as follows: Let X_t, Y_t time series and let $X_t^{(d,\tau)}, Y_t^{(d,\tau)}$ be embedding vectors belonging to the reconstructed manifolds M_X, M_Y respectively. A cross-mapped estimate of Y_t based on M_X is constructed by considering the value X_t and locating its $d+1$ nearest neighbors on M_X. Denote the time indices of the neighbors (from nearest to farthest) by $t_1, t_2, ..., t_{d+1}$. The time series Y_t is subsequently estimated based on these indices via a locally weighted average of the following form:

$$\widehat{Y}_t|M_X = \sum_{i=1}^{d+1} w_i Y_{t_i} \tag{6.10}$$

Using the analogous definition, cross-mapped estimates of X_t based on M_Y are obtained. Finally, the cross-mappings are evaluated via the Pearson correlation coefficient between Y_t and $\widehat{Y}_t|M_X$, and similarly between X_t and $\widehat{X}_t|M_Y$. In the first case, high correlation indicates that system X is causing system Y (i.e. time series X_t is causing time series Y_t) while in the second case the opposite applies.

In the original paper, inference on causality happens by visually inspecting the values of the correlation coefficient, which (in case of causation) should converge to values close to 1, as the length of the time series increases. This visual assessment does not allow for automation of the procedure that a benchmark study requires, so other researchers developed significance tests for the correlation values found based on surrogate time series creation (see Section 3.5), which we also use here.

Experimental Design

In terms of parameter selection for its application, CCM is a particularly simple method, as the embedding dimension d and delay τ are the only important choices. An embedding dimension of 3 was used, while the embedding delay was set to 1. As a bivariate method, CCM was separately applied to each directed variable pair of the dataset impacting its computational complexity.

Qualitative Properties

CCM focuses on detection of causal interactions instead of causal delays, however a follow-up paper Ye et al. (2015) incorporates (even instantaneous) causal delays in the method. Self-causation and polyadic relations are not inferred, and as a bivariate method CCM may be challenged by the

existence of (un)observed confounders within a dataset. Particular caution is required when a strong uni-directional causal relation $X \to Y$ exists in a dataset as CCM might incorrectly infer the opposite relation $Y \to X$ too Sugihara et al. (2012). As a nearest-neighbor based method, CCM is robust in detecting non-linear relations, featuring discrete or continuous data.

6.5.5 MVGC

Multivariate Granger Causality is the implementation of Wiener's idea for causality by Granger. This was presented in Section 2.2.1, albeit for the univariate case. For the purposes of a multivariate dataset of time series, a multivariate adaptation is needed. This extension is obtained by considering the multivariate analog of auto-regressive modelling shown in Section 2.2.1, that was also encountered in Section 4.1.1: Vector Auto-Regressive (VAR) modelling Lütkepohl (2005).

Method Presentation

Given a multivariate time series, a VAR model of order p has the following form:

$$U_t = \sum_{k=1}^{p} A_k U_{t-k} + \varepsilon_t \tag{6.11}$$

where U_t is stationary, A_k is a matrix containing the model coefficients and residuals ε_t are vectors. The matrix A_k alongside the residual covariance matrix Σ are the model parameters. Important for VAR(p) theory is the autocovariance function Γ of U_t (see Definition 4.1.2 for the bivariate case) that relates to the parameters of the VAR model through the *Yule-Walker* equations:

$$\Gamma_k = \sum_{\ell=1}^{p} A_\ell \Gamma_{k-\ell} + \delta_{k,0}\Sigma \quad , k \in \mathbb{Z} \tag{6.12}$$

Suppose that the multivariate time series U_t is split into two jointly distributed time series X_t, Y_t, i.e $U_t = [X_t, Y_t]^T$. The VAR(p) model is then decomposed:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \sum_{k=1}^{p} \begin{bmatrix} A_{xx,k} & A_{xy,k} \\ A_{yx,k} & A_{yy,k} \end{bmatrix} \begin{bmatrix} X_{t-k} \\ Y_{t-k} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix} \tag{6.13}$$

$$\Sigma = Cov(\begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix}) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{6.14}$$

To examine the causal relation $Y \to X$ the first component X_t of the decomposition is considered:

$$X_t = \sum_{k=1}^{p} A_{xx,k} X_{t-k} + \sum_{k=1}^{p} A_{xy,k} Y_{t-k} + \varepsilon_{x,t} \tag{6.15}$$

Similarly to the univariate case presented before, Y is causing X if the null hypothesis of no causality is rejected:

$$H_0 : A_{xy,1} = ... = A_{xy,p} = 0 \tag{6.16}$$

This hypothesis is tested via the multidimensional analogue of (2.32) that is

$$\mathcal{F}_{Y \to X} = \log \frac{\det \Sigma'_{xx}}{\det \Sigma_{xx}} \tag{6.17}$$

where Σ_{xx} is the covariance matrix of the residual vector $\varepsilon_{x,t}$ of the full model (6.15) while Σ'_{xx} is the covariance matrix of the residual vector of the nested model that assumes H_0. Extensions to the conditional case $Y \to X|Z$ are parallel to the univariate case, and results on the distribution of $\mathcal{F}$ still hold.

Experimental Design

In the specific implementation used, the user first specifies a maximum model order $p_{\max}$, and for each $\{1, ..., p_{\max}\}$ a VAR model is fitted into the data. A specific model order p is then selected based on information criteria (AIC or BIC). Subsequently, the corresponding VAR model parameters A_k, Σ are estimated using e.g. OLS, and the autocovariance function Γ is obtained from them using the Yule-Walker equations. Then, for each causal relation $Y \to X|Z$ to be examined, the corresponding VAR parameters for both the full model (6.15) and its nested alternative are calculated using the autocovariance function and the Yule-Walker equations. An estimate for $\mathcal{F}_{Y \to X}$ is therefore obtained, and its significance is assessed via its (known) distribution.

For this benchmark study $p_{\max} = 10$ was used, and p was selected based on the AIC. The estimation of the model parameters was done with OLS. All directed pairs $X \to Y|Z$ were investigated for significance, where Z consisted of all other time series of the dataset. The F statistic (6.17) was used for significance of estimates, at the author's default significance level of 0.01 while correcting for multiple testing using the false discovery rate Benjamini and Hochberg (1995).

Qualitative Properties

MVGC does not accompany every causal relation detected with an explicit causal delay, however the model order can provide information on the maximum causal delay found. It does not investigate self-causation, nor does it infer instantaneous causality (for instantaneous GC see Kirchgässner and Wolters (2007)). MVGC can filter out confounding effects (provided they are observed, see Guo et al. (2008) for unobserved confounders) and it can also account for indirect causation. MVGC investigates dyadic relations. As Barnett and Seth (2014) note, even if the VAR models used to formalize MVGC are linear, a general class of stationary processes (including many non-linear ones) can be modelled with them provided the model order is large enough Anderson (1971). In reality however, to avoid overfitting a dataset, the model order will remain relatively small effectively restraining the ability of MVGC to capture non-linearities. MVGC can handle discrete or continuous data. Stationarity is assumed, and MVGC automatically tests for it.

6.5.6 TCDF

The Temporal Causal Discovery Framework was recently introduced in Nauta et al. (2019). It approaches causality from a neural-network standpoint. While multiple neural-network based methods already exist broadly in causal inference Goudet et al. (2017), Kalainathan et al. (2018), Rossi et al. (2020), TCDF is the first method designated for temporal data.

Method Presentation

Given N time series $(X_t^{(1)}, ..., X_t^{(N)})$, TCDF performs causal discovery in steps. Initially, N independent Convolutional Neural Networks (CNNs, Goodfellow et al. (2016)) are defined, where the j-th network is independently trained to predict its target time series $X_t^{(j)}$. In doing so, it uses the whole dataset (including $X_t^{(j)}$). TCDF employs a specific class of CNNs called *attention-based* CNNs Yin et al. (2015): these networks feature an attention mechanism that explains on which data network j *attends to* when predicting its target time series. Therefore, if a network j attends to time series $X_t^{(i)}$ for predicting $X_t^{(j)}$, the method infers that $X_t^{(i)}$ is a potential cause of $X_t^{(j)}$ (closely resembling the original idea of Granger causality). Subsequently, a causal validation step is performed. During this step, TCDF relies on intervening in the dataset in order to derive true causalities. For each potential cause $X_t^{(i)}$ of time series $X_t^{(j)}$ an *intervened* dataset is created, equal to the original besides the values of $X_t^{(i)}$ that are permuted. The same network j is re-used to obtain a new prediction of $X_t^{(j)}$. TCDF then uses a specialized routine to compare the predictive performance between the two cases and infer on whether $X_t^{(i)}$ is a true cause of $X_t^{(j)}$.

Experimental Design

The TCDF algorithm was applied with 1000 epochs for each network trained and kernel size 4 dictating a maximum temporal search for causal interactions of 3 lags. No hidden layers were used in the depthwise convolution, while the rest of the parameters pertaining to training the CNNs were defaulted to the values recommended by the authors.

Qualitatitve Properties

TCDF supplements the final step of causal validation with delay discovery. Since for the prediction of any time series, the whole dataset is used (including the time series itself) it captures self-causation. TCDF is able to discern instantaneous causations, and it can also deal with confounders, being able to also hypothesize their existence if they are not observed. TCDF does not support the detection of polyadic relations. TCDF can detect non-linear relations on discrete, continuous and even non-stationary data.

6.5.7 PDC

Partial Directed Coherence is an extension of Granger causality to the frequency domain. As such, it shares certain properties with GC, and it is also formalized through linear VAR models. Non-linear variations of PDC were recently proposed (e.g. He et al. (2014)), however in this study we focus on the traditional (linear) PDC, as it was introduced in Baccalá and Sameshima (2001), and later extended to generalized PDC in Baccala et al. (2007). Frequency domain causality measures are very popular in the analysis of neurophysiological signals Pereda et al. (2005), Kindlmann and Burel (2008), Sakkalis (2011) and benchmark studies of frequency domain measures from the perspective of neurosciences exist Wang et al. (2014), Haufe et al. (2013), Sommariva et al. (2017).

Method Presentation

Given K time series $X_t = (X_t^{(1)}, ..., X_t^{(K)})$, we consider a VAR model of order p (as in the case of MVGC). So, for time series $X_t^{(j)}$ we have the following model:

$$X_t^{(j)} = \sum_{k=1}^{K} \left(a_{k,1}^{(j)} X_{t-1}^{(k)} + ... + a_{k,p}^{(j)} X_{t-p}^{(k)} \right) + \varepsilon_t^{(j)}, \quad j = 1, ..., K \tag{6.18}$$

As for MVGC, the coefficients $a_{i,r}^{(j)}$ may be estimated by standard approaches (e.g. OLS). Then, we consider their frequency transform:

$$A_i^{(j)}(f) = \begin{cases} 1 - \sum_{r=1}^{p} a_{i,r}^{(j)} e^{-2\pi i f r} & \text{if } i = j \\ - \sum_{r=1}^{p} a_{i,r}^{(j)} e^{-2\pi i f r} & \text{if } i \neq j \end{cases} \tag{6.19}$$

where the frequencies f are discrete. The generalized partial directed coherence at frequency f from time series X_i to X_j is

$$GPDC_{X_i \to X_j | Z}(f) = \frac{\frac{1}{\sigma_{i,i}} |A_i^{(j)}(f)|}{\sqrt{\sum_{k=1}^{K} \frac{1}{\sigma_{k,k}^2} |A_i^{(k)}(f)|^2}} \tag{6.20}$$

GPDC quantifies the effect from X_i to X_j at frequency f normalized by the effect of X_i to every other variable. Here, $\sigma_{k,k}^2$ is the variance of $X_t^{(k)}$ and $Z = X \setminus \{X_i, X_j\}$. Diverging from the methods presented before, GPDC is defined as a function of frequency. In practice, researchers calculate the average GPDC over a set of frequencies (frequency band), and domain knowledge

plays a role in the selection of the band for an analysis. Similarly to TE, the null hypothesis of zero GPDC and therefore no causation is tested via the creation of surrogate time series.

Experimental Design

For each time series pair X_i, X_j the GPDC is estimated and tested for significance (at a significance level $a = 0.01$), forming a binary causal adjacency matrix. The sampling frequency is 1, sampling 512 points in low-dimensional datasets, and 128 points in high-dimensional datasets due to computational reasons. The maximum model order is restricted to 4 in the high-dimensional case and to 10 in the low-dimensional case. AIC is used for model selection, while the Nuttall-Strand algorithm Schlögl (2006) is used for VAR estimation. At every frequency, the binary causal adjacency matrix is derived. We average these matrices over all frequencies, and round up to 1 all elements of the averaged adjacency matrix that are bigger than 0.5, while all other elements are set to 0.

Qualitative Properties

PDC does not function on the time domain, and no causal delays are reported. It is able to infer self-causal relations. The specific implementation used is not able to infer instantaneous causation on the frequency domain, however an extension provided in Faes et al. (2013a) can fulfill this task. PDC is able to account for confounders and distinguish direct from indirect causation, provided they are observed. Relations inferred are dyadic. Equivalently to MVGC, VAR modelling can be robust to non-linear patterns but this may be hindered by the model selection procedure. PDC assumes stationary time series, featuring discrete or continuous data.

6.5.8 Summary of qualitative properties

Based on the list of important qualitative properties for methods performing causal inference in time series we compiled and on whether each method investigated fulfills these properties discussed in this chapter the following table is created. It summarizes the properties of all methods and proposes a classification scheme for causal inference methods in time series. Properties are reported with respect to the specific implementation used in this benchmark.

Method	MTE	PMIME	PCMCI	MVGC	TCDF	PDC	CCM
Delay discovery	✓	✗	✓	✗	✓	✗	****
Self-causation	✗	✗	✓	✗	✓	✓	✗
Instantaneous causality	✓	✗	✓	✗	✓	***	****
Observed confounders	✓	✓	✓	✓	✓	✓	✗
Unobserved confounders	✗	✗	✗	✗	**	✗	✗
Polyadic relations	✗	✗	✗	✗	✗	✗	✗
Non-linear data	✓	✓	✓	*	✓	*	✓
Non-stationary data	✗	✗	✗	✗	✓	✗	✗
Bivariate/Multivariate data	both	both	both	both	both	both	bivariate
Discrete/Continuous data	both	both	both	both	both	both	both

Table 6.2: Overview of properties for each method examined. *: Barnett and Seth (2014). **: Nauta et al. (2019). ***: Faes et al. (2013a).****: Ye et al. (2015).

Chapter 7

Results

The results of the benchmark study conducted in Chapter 6 are listed here. The methods are evaluated quantitatively, as described in Section 6.4.2. Moreover, their performances are compared, visualized and remarks on the overall results are made. All methods were executed on an 8$^{\text{th}}$ generation Intel® Core© i5-8365U CPU.

7.1 Method performance

We apply every method presented in Chapter 6 to the datasets introduced in Section 6.2 utilizing the experimental design detailed for each method in its corresponding section. To elaborate on the evaluation of each method in practice, we present the full results of the data group H_1 for multivariate transfer entropy. Iterating MTE over 5 such datasets and comparing the estimated causal graph with the ground truth, we obtain the results listed in Table 7.1.

Category H_1	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
True positives	30	30	36	26	36	
False positives	9	8	4	23	7	
True negatives	335	336	340	321	337	
False negatives	6	6	0	10	0	
Sensitivity	0.83	0.83	1	0.72	1	0.88
Specificity	0.97	0.98	0.99	0.93	0.98	0.97
F1 score	0.80	0.81	0.95	0.61	0.91	0.82
MCC	0.78	0.79	0.94	0.57	0.91	**0.80**
Time (in seconds)	10845	10395	11097	11811	11460	**11097**

Table 7.1: Full MTE results on the first data category, rounded to two decimals. The average MCC and the median runtime are both highlighted with bold.

This table contains all relevant performance evaluation metrics introduced in the previous chapter for this specific method / data category combination. In the next section, for the simulated Hénon datasets we will only present the average MCC and runtime of each method and shortly comment on its performance. For the real dataset, the F1 score will be reported instead of the MCC, as the latter was found to be too strict for the evaluation of a very low dimensional dataset. The results will be subsequently summarized in Table 7.9. The next section includes a series of visualizations of the results, their interpretation, and insights gained from them. Here, results are reported and discussed following the order the methods were presented.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.80	0.97	0.80	0.94	0.88	0.80 (F1 score)
Time (median, seconds)	11097	795	10993	803	5922	-

Table 7.2: Summary results for MTE.

MTE

Multivariate TE performs very well on the low dimensional H_2 and H_4 data groups, and its performance as well as its running time appear to be significantly impacted on higher dimensional datasets. It is the slowest method with the average median running time over the different data groups exceeding 1.5 hours. MTE performs equally great in the low dimensional loosely coupled group (H_4) and in the low dimensional strongly coupled one (H_2). The robustness of TE in low dimensions may be attributed to the fact that this particular implementation uses the first KSG estimator for TE (3.20). This estimator bases the threshold for counting points in the marginal spaces on the distance of every point to its k^{th} nearest neighbor on the joint space, and not marginally. As it is noted in Kraskov et al. (2003), for low dimensional data this will not be harmful; only for high dimensional data will the second KSG estimator perform better. Noting the significant difference in MTE performance between low and high dimensional datasets, this remark should be taken into account in TE analyses. On real data, MTE also performed well.

PMIME

PMIME performed perfectly in all Hénon data, precisely retrieving the correct causal graph at every iteration. Contrasting the results between the two information theory methods studied so far, PMIME outperformed MTE both in terms of performance and speed. Indeed, a staggering difference is noted in the computational complexities of the two methods. As remarked in Kugiumtzis (2013) the fact that PMIME bypasses the computationally exhausting step of significance testing of estimates dramatically improves its speed compared to MTE. PMIME was however challenged by the real dataset, registering a below average result.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	1	1	1	1	1	0.50 (F1 score)
Time (median, seconds)	2075	127	2078	126	1101	-

Table 7.3: Summary results for PMIME.

PCMCI

PCMCI performed decently well on simulated data, and was among the top performing methods in the real dataset. The main asset of this method is its computational speed, as PCMCI was the fastest among the methods examined. An interesting observation regarding PCMCI is the fact that it seems to actually benefit from weakly coupled data in terms of performance.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.67	0.60	0.78	0.85	0.72	0.80 (F1 score)
Time (median, seconds)	25	1	21	1	12	-

Table 7.4: Summary results for PCMCI.

MVGC

MVGC on average performed relatively well, however its performance was inconsistent and highly impacted by different coupling strengths and dimensionalities; MVGC significantly benefited from high dimensional data. On real data, MVGC showcased above-average performance.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.84	0.29	0.72	0.53	0.70	0.67 (F1 score)
Time (median, seconds)	383	3	220	3	152	-

Table 7.5: Summary results for MVGC.

TCDF

On simulated data, TCDF also performed decently well. It attained a balance between consistent performance over different configurations, computational speed (scaling favorably as the number of variables increased) and general robustness in the data it can accommodate, being suitable even for non-stationary data. On the other hand, TCDF performed badly on the real dataset, essentially failing to detect causality.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.83	0.76	0.65	0.70	0.74	0 (F1 score)
Time (median, seconds)	73	17	73	17	45	-

Table 7.6: Summary results for TCDF.

PDC

On simulated datasets, PDC was the best non-information theoretic method. It exhibited consistently high performance throughout the different categories. The specific implementation used was found to be significantly slower as the number of time series increased, which should be attributed to inefficient coding routines. On real data, PDC displayed mediocre performance.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.85	0.86	0.86	0.82	0.85	0.56 (F1 score)
Time (median, seconds)	4936	14	4866	13	2457	-

Table 7.7: Summary results for PDC.

CCM

CCM was the worst performing method in the Hénon map datasets. This may be attributed to the fact that CCM is only able to make bivariate inferences. As a result, it is the only method that does not account for the effects of the other variables on each causal relation it investigates. In the low dimensional real dataset consisting of 3 variables only, CCM performed well.

An overview of the results discussed so far is included in Table 7.9. In this table, methods are sorted based on their overall average performance (on simulated data) from the best performing (top) to the worst performing (bottom) methods. The results presented so far are subsequently discussed and visualized in the next section, and methods are comprehensively compared.

Category	H_1	H_2	H_3	H_4	Avg. Hénon	Real Data
MCC (average)	0.36	0.41	0.28	0.44	0.37	0.80 (F1 score)
Time (median, seconds)	621	31	587	31	318	-

Table 7.8: Summary results for CCM.

Category	H_1	H_2	H_3	H_4	Average	Runtime	Real Data (F1 score)
PMIME	1	1	1	1	1	1101	0.50
MTE	0.80	0.97	0.80	0.94	0.88	5922	0.80
PDC	0.85	0.86	0.86	0.82	0.85	2457	0.56
TCDF	0.83	0.76	0.65	0.70	0.74	45	0
PCMCI	0.67	0.60	0.78	0.85	0.72	12	0.80
MVGC	0.84	0.29	0.72	0.53	0.70	152	0.67
CCM	0.36	0.41	0.28	0.44	0.37	318	0.80

Table 7.9: Summary of all results.

7.2 Visualizations and insights

In this section, the simulated and real data are going to be treated separately. We begin with visualizing different quantities pertaining to performance and runtime of methods in simulated data.

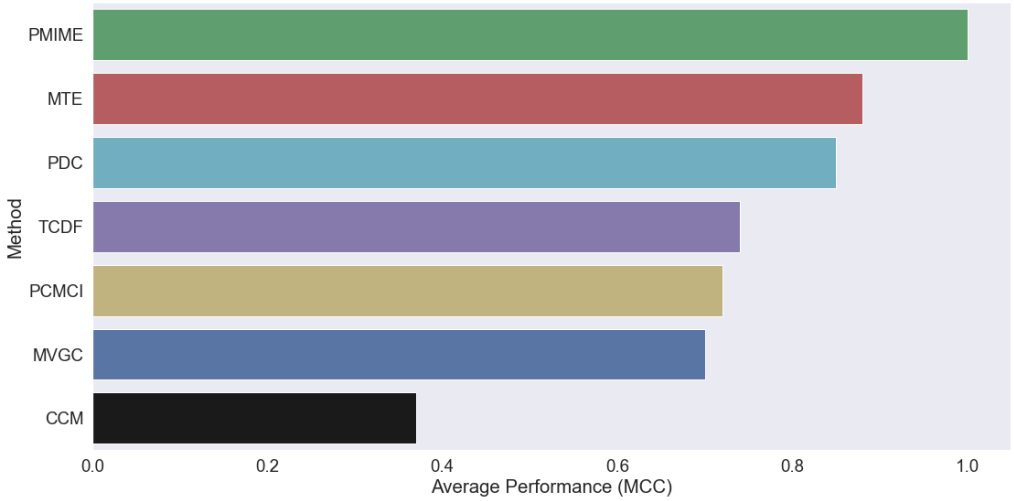

Figure 7.1: The (overall) average column of Table 7.9 visualized per method.

First, Figure 7.1 contains the overall performance of each method, visualizing how the values listed in Table 7.9 compare. We notice that the information theoretic methods register the best performance. Both are based on mutual information: they do not make any assumptions in modelling the relations between variables and admittedly benefit from it. PDC is the best among the rest of the methods, and as it compares very favorably to MVGC, it demonstrates the utility of a frequency domain transform preceding causal inferences from a VAR model. TCDF maintains good performance and offers great robustness, while PCMCI (alongside the linear conditional independence tests used in this project) is an ideal candidate for a preliminary analysis of any dataset, partially due to its speed. More research and a multivariate adaptation accounting for interactions is required to transform CCM into a good fit for the task this project undertook.

While the barplot above allows for a simple and direct comparison between methods, it is solely based on averages. A boxplot of all MCC values per method is thus visualized in Figure 7.2, aiding us in obtaining a more complete view of how the methods performed.

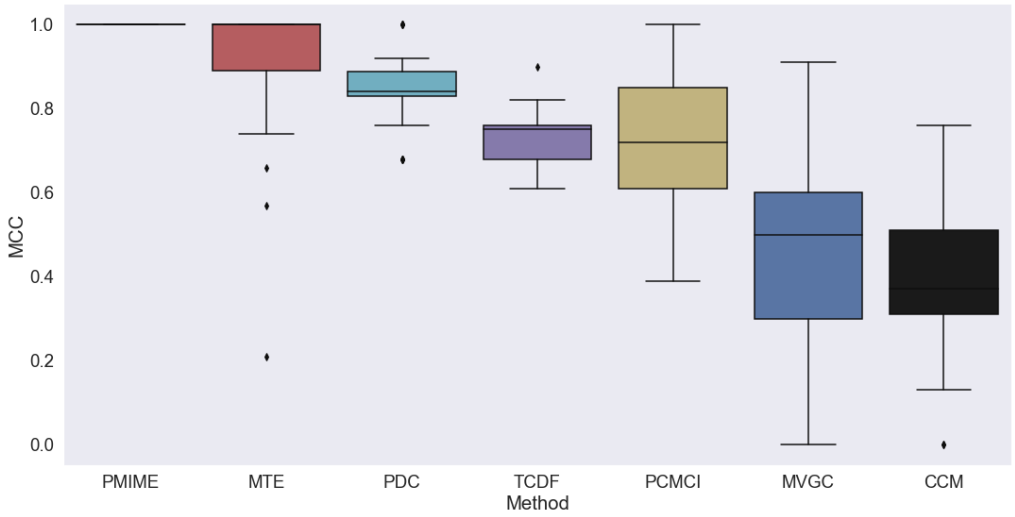

Figure 7.2: Boxplots showing the performance dispersion of methods throughout different iterations of the benchmark.

Notice that the dispersion in the performance of each method varies. A consistent method showcasing less variant performance such as PDC or TCDF might be preferential to a faster (PCMCI) or even better performing (MTE) method that is less stable or exhibits outlying performances.

In the presentation of results so far, we have aggregated performances over all 4 Hénon map data categories. We can also utilize the differences between the 4 different data configurations, to arrive at more specialized insights; the respective plots are included in Appendix B.

We proceed with a short discussion of running times of methods. Figure 7.3 contains a barplot visualizing the average median runtime over the 4 Hénon data categories per method on a logarithmic scale. An additional graph pertaining to the effect an increase in dimensionality of a dataset has on running times of each method is included in Appendix B.

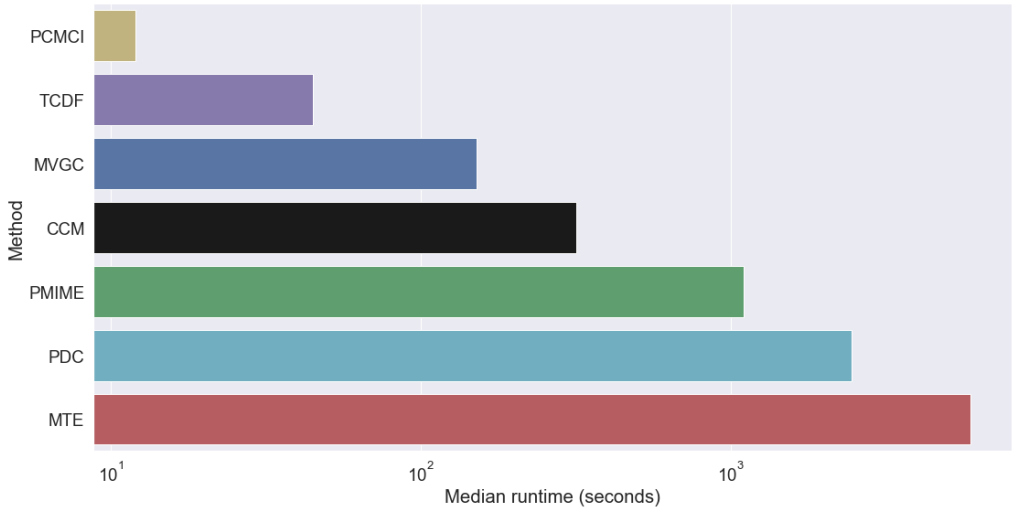

Figure 7.3: Barplot of the average median running time of an iteration of each method (log scale).

Note that the two information theoretic methods are among the slower methods, with MTE being the slowest. In addition, as PDC was also found to be very slow with the average median PDC iteration taking thousands of seconds, we observe a reversal of Figure 7.1 that featured the performances of methods. The following scatterplot is very informative regarding the trade-off between computational complexity and performance.

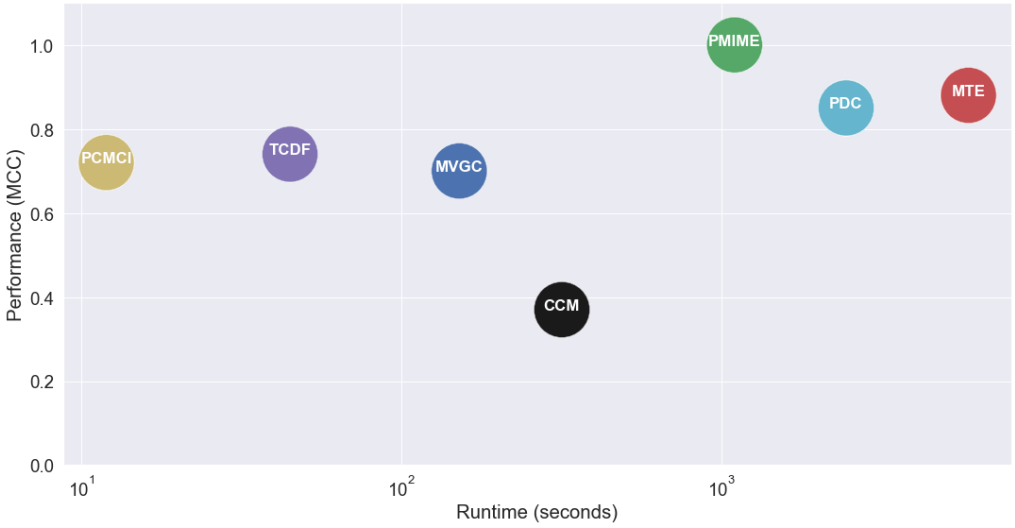

Figure 7.4: Scatterplot visualizing the trade-off between method speed and method performance.

According to the findings of this benchmark study we conclude that, when selecting a causal inference method for the analysis of time series, speed against performance constitutes a significant dilemma. Model-based approaches (PCMCI, GC, TCDF) are able to attain relatively good performances in discovering the causal structure of a temporal dataset - and the assumptions they make to do so may not necessarily hamper their versatility (TCDF, VAR modelling). However, perfect or near-perfect results may not be expected; if performance is the main priority, the model-free framework provided by information theory will, generally, provide a better fit. Depending on the context of an application and on the end goal of a causal inference study, it may or may not be beneficial to sacrifice performance for speed.

Concluding the chapter, we showcase the results obtained from the real dataset. Figure 7.5 contains a barplot visualizing the performance of the methods in real data.

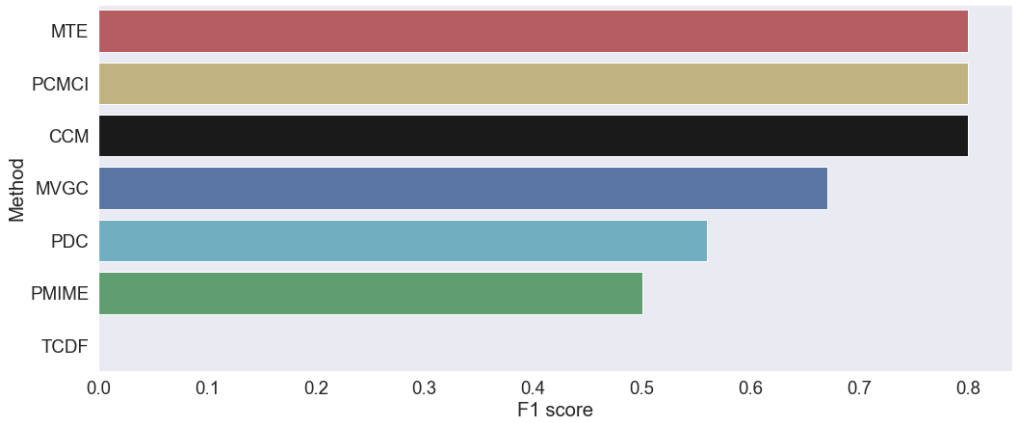

Figure 7.5: Barplot visualizing the performance (F1 score) of each method on the real dataset.

It should be noted that the best performing methods all retrieve a fully connected causal graph. In this 3-dimensional dataset, this will translate to a good F1 score however the inability methods generally showcased in detecting a true negative (i.e. lack of causation) might be indicative of results that are not robust. Moreover, the unobserved confounder(s) that exist combined with the very low number of time series and the relatively low amount of datapoints used for causal discovery, constitute a tough case study.

Chapter 8

Conclusions

8.1 Summary and conclusions

This thesis investigated two research questions. The first pertained to the study of transfer entropy from a non-stationary perspective. The second concerned the development of a framework for the comprehensive and objective comparison between causal inference methods in time series data.

Transfer entropy and non-stationarity

A concrete non-stationary system was introduced and studied in detail, deriving probabilistic results and obtaining a theoretical expression for transfer entropy in it. Doing so, it extended the results featured in the paper it was based on, and possibly provided the first exact results for transfer entropy in a non-stationary system in literature.

This task culminated in the examination of a non-stationary estimator for transfer entropy, that was subsequently evaluated and compared to the theoretical transfer entropy values. This estimator was introduced in a recent paper for differential entropy, and for the purposes of the project was successfully adapted and implemented for the transfer entropy case. A trivial extension of the system was also considered under which the transfer entropy results proved to remain invariant.

The asymptotic behavior of transfer entropy was also studied and its convergence was proved. As convergence of transfer entropy was found to generally happen fast, sensitivity analysis was performed in its limit. Due to the specific structure of the system that was studied, a connection of the results to the concept of a sensor measuring a physical quantity was illustrated. This enabled the examination of the impact of the characteristics of a sensor on the magnitude of the information flow it receives from the physical process it aims to measure.

Benchmark framework for causal inference methods

After consulting the relevant literature and minding the specific context of ASML, a list of 7 methods coming from diverse backgrounds and featuring heterogeneous characteristics was proposed to be included in the study.

In parallel, an appropriate dataset was selected for simulation, and its selection was motivated by briefly reviewing its theory. The characteristics of this dataset were varied and multiple different yet identical in their macroscopic structure datasets were generated to evaluate the selected methods on. Supplementing the simulated data, a suitable real dataset was also used in the analysis.

A list of important qualitative properties to be reported for each method was compiled, motivated by the particular challenges real data might pose, and the procedure for quantifying the performance of each method in the data was substantiated. The methods were serially applied on the same datasets, and their performance and running time were obtained.

The results of the benchmark framework were pooled in a single dataset that was used to retrieve informative visualizations illuminating and enabling discussion regarding how the methods compared, the dispersion of the performance of each method and the proficiency methods showcased under data with varying characteristics. Insights on the computational complexity of the methods as well as on the balance each method attains between speed and performance were also retrieved.

Overall conclusion

With respect to the first research question, the main goal of this thesis was the estimation of transfer entropy in non-stationary time series. As the direct estimation of transfer entropy in the general non-stationary case without additional hypotheses remains an open question, two assumptions were made that impacted the generality of results: first, a concrete system was investigated, and second, it was proved to belong to the smaller class of non-stationary time series with stationary increments. In this context, the estimator that was examined showcased relatively good performance in estimating transfer entropy; although further improvements in terms of bias correction are required. Being aware of the limitations discussed, we conclude that this estimator can be considered as an option for the estimation of more involved datasets - provided stationarity of increments holds.

Regarding the second research question, the findings of the benchmark study seem to uphold the thesis that the model-free framework of information theory should be used if the performance of causal inference methods is the central concern. The findings simultaneously demonstrate that aiming for perfectly performing methods comes at a significant cost in terms of computational complexity; a balance between the two should be attained. The qualitative classification scheme proposed for causal inference methods and the heterogeneity observed in their properties, elucidate the importance of carefully considering the context and the data involved in a causal inference study before embarking on it.

8.2 Discussion and recommendations

This section contains remarks related to important details of the project. Potential shortcomings of the approach taken are also mentioned here.

Transfer entropy and non-stationarity

Estimation of entropy terms: Based on the advice of Granero-Belinchón et al. (2019), the TE estimator studied in Chapter 4, essentially utilizes a box-kernel density estimator for each entropy term. While in practice it performed decently, a more thorough investigation of this estimator is required; e.g. it has not been compared to other kernel density estimators, or other entropy estimators in general.

Method of estimation: The estimator investigated is derived by applying the estimator of Granero-Belinchón et al. (2019) in different entropy terms appearing on the definition of transfer entropy. In mutual information estimation, this technique is known to adversely impact the overall estimate Kraskov et al. (2003) as separate biases may accumulate. Despite not proving detrimental in this study, this might generally harm the estimates.

Low embedding dimension: The system studied in Chapter 4 features a very short time dependency on its past: only the present and immediate past states play a role. This allowed us to study a small embedding dimension in Section 4.3. The estimator might be negatively affected as larger embedding dimensions are considered.

Pattern of non-stationarity found: Due to the simplicity of the structure of the system studied and the assumptions made, we formally proved that the cTE converges in time. This is as much of an insight as a limitation: the pattern of non-stationarity found is rather trivial

(and quickly convergent to a simple logarithmic function). An example of a more involved non-stationary theoretical derivation for TE remains elusive.

Time index: Throughout the thesis, time is discrete. Minding the difference in time scales over which data are registered within ASML and instantaneous causation that might arise from this feature, continuous time TE might provide a better fit - although more theoretical developments are required.

Benchmark framework for causal inference methods

Unobserved confounders: The presence of unobserved confounders is ubiquitous in real datasets, and methods showcasing robustness to them are advantageous. While the real dataset used in this project was influenced by hidden confounders, the simulated dataset did not have any. A hidden confounder can be conveniently simulated by simply dropping a confounding time series from the simulated dataset, and observing the difference in method performance could be insightful.

Causal delays: The causal delays reported by a method are not considered in their evaluation. This is to ensure a homogenized comparison, as some methods do not detect causal delays. By appropriately extending the evaluation metric used, or by selecting a different one, this information may also be taken into account, providing a more complete picture of how the methods performed and compared.

Computational restrictions I: Due to restrictions in the computational power available, methods were iterated a moderate amount of times over a few different data configurations. It is usually the case in literature to e.g. iterate methods not over 2 different coupling strengths (low/high) but over 10's of different coupling strengths, hence obtaining a more precise overview of their performance.

Computational restrictions II: Limited computational power also restricted the different kinds of datasets included in the analysis. The Hénon map is but one of multiple frequently used datasets for such a purpose. Coupled Lorenz systems, or data simulated from VAR models could have also been part of the study. Recent relevant work Finn and Lizier (2020) incorporates different network structure motifs into similar studies as well.

8.3 Potential of causal inference within ASML

An important goal within ASML is the performance of predictive maintenance in lithography systems. A prerequisite goal however, is the development of a reliable data-driven method that performs fault diagnosis and root cause analysis – in (close to) real time. A first step towards this method is understanding the complex interactions happening in the system, i.e. unveiling its structure.

Causal inference is a suitable fit for this task. As a data-driven field, a causal inference method might reveal a non-trivial interaction occurring in a system that was not previously found due to the complexity of the system's interactions. Such causal insights reach beyond questions on associations and standard statistical or machine learning methods may not be compatible with such a task. This project studied a variety of relevant methods, exploring very diverse approaches and illustrated their utility.

Transfer entropy in particular has been successfully applied within ASML in the past. TE requires minimal domain knowledge in order to analyze a dataset, and its results can be conveniently visualized and communicated with domain experts. As this thesis also demonstrated, a major drawback for TE is its computational complexity; this may pose a challenge in its application within a production environment. Nevertheless, TE and its different versions are flexible, and attempts can be made to relieve the computational load that TE carries. A relevant point is made on the next and final section.

8.4 Future research

There are several research questions that can be pursued as a continuation of this project:

The embedding dimensions selected for TE estimation are very important as they impact the estimate and alter the computational load. Ideally, the embedding dimension of the target signal should be optimized such that self-prediction is maximized, to separate information storage from information transfer Lizier et al. (2008). Investigating how embedding dimensions impact TE estimation results is therefore an interesting question.

The high computational demand of TE depends on both the sample size and especially on dimensionality. Therefore, reducing the size of the dataset while retaining as much information as possible can be greatly beneficial. In terms of reducing dimensionality, a variety of well know methods from literature could be tried, such as PCA Jolliffe and Cadima (2016) or Isomap Tenenbaum (2000).

Once a causal graph has been established using a causal inference method, a natural next step is to quantify a notion of how "influential" each node is in the graph. Some recent papers Streicher and Sandrock (2019), Murin et al. (2018) have proposed using centrality measures from graph theory to quantify this (although within structural causal models caution is required Dablander and Hinne (2019)). On the other hand, centrality measures have been found to correlate with the firing rate of a neuron Fletcher and Wennekers (2018), rendering them important in this specific case. This research direction might be fruitful; in this context, controlling for edge multiplicity might be important.

Considering the time-dynamic context, the above idea opens up more possibilities. Given a multivariate time series dataset over the same time scale, multiple consecutive time windows may be causally analyzed with any method presented here. This would result in a sequence of directed graphs being derived from the dataset - and the influence metric of each node would be a time series itself. These time series may be subsequently analyzed, e.g. using standard time series analysis techniques Hyndman and Athanasopoulos (2018), or even clustered, using e.g. dynamic time warping Müller (2007). Clustering similarly-behaving nodes (i.e. variables of the system) can be promising in revealing underlying sub-systems, that may have been invisible otherwise - simplifying the network.

This sequence of directed graphs can be further explored. Anomaly detection in dynamic networks Ranshous et al. (2015) may be performed on it, interestingly combining causal inference (that yields a directed graph) with the time-dynamic context (graph changing over time) and the end goal (prognostics/diagnostics). For this direction to be sensible, a high-quality causality method is required. The current project extensively researched that. The time-varying causal evolution of a time series dataset was studied in Jiang et al. (2017).

With regards to non-stationary TE estimation, an alternative approach that was not considered in this work may be based on Hegger et al. (2000). In this paper, the authors argue that if a time series is non-stationary because of a slow drift, it may suffice for its analysis to *over-embed* it, i.e. appropriately increase the embedding dimension.

The concept of *copula entropy* Nelsen (2007), was shown to be the same with mutual information Ma and Sun (2008). Since then, it has been used in relation to Granger Causality Hu and Liang (2014), and more recently for TE estimation Jian (2019) as well as in applications Hao and Singh (2015), Sun et al. (2019). Further researching copula entropy in relation to TE could provide another research direction.

In this project, exact results for transfer entropy were derived in a random walk system. Aspiring to generalize these results, a time series decomposition method that contains a random walk as one of its components may be useful. A suitable fit might be provided by the *Beveridge-Nelson* decomposition Beveridge and Nelson (1981) that decomposes a non-stationary ARIMA process to a stationary time series and a random walk.

Bibliography

János Aczél and Zoltán Daróczy. On measures of information and their characterizations. *New York*, 122, 1975. 98

Theodore W Anderson. *The statistical analysis of time series*, volume 19. John Wiley & Sons, 1971. 72

Andras Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19(3-4):163–193, 2001. doi: 10.1002/rsa.10019. 23, 24

Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(01):17–41, 2008. doi: 10.1142/s0219525908001465. 15, 59

Luiz A. Baccalá and Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84(6):463–474, 2001. doi: 10.1007/pl00007990. 73

Luiz A. Baccala, K. Sameshima, and D.Y. Takahashi. Generalized partial directed coherence. In *2007 15th International Conference on Digital Signal Processing*. IEEE, 2007. doi: 10.1109/icdsp.2007.4288544. 62, 73

G. Baier and M. Klein. Maximum hyperchaos in generalized hénon maps. *Physics Letters A*, 151 (6-7):281–284, 1990. doi: 10.1016/0375-9601(90)90283-t. 54

Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88(17), 2002. doi: 10.1103/physrevlett.88.174102. 30

Lionel Barnett and Terry Bossomaier. Transfer entropy as a log-likelihood ratio. *arXiv preprint*, 2012. doi: 10.1103/PhysRevLett.109.138105. 14, 33

Lionel Barnett and Anil K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, 2014. doi: 10.1016/j.jneumeth.2013.10.018. xi, 62, 72, 74

Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103(23), 2009. doi: 10.1103/physrevlett.103.238701. 14

Margret Bauer, John W. Cox, Michelle H. Caveness, James J. Downs, and Nina F. Thornhill. Finding the direction of disturbance propagation in a chemical process using transfer entropy. *IEEE Transactions on Control Systems Technology*, 15(1):12–21, 2007. doi: 10.1109/tcst.2006.883234. 11

J. Beirlant, E. Dudewicz, L. Gyor, and E.C Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6, 1997. 23

Michael Benedicks and Lennart Carleson. The dynamics of the Henon map. *Annals of Mathematics*, 133(1):73–169, 1991. ISSN 0003486X. 54

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 72

Stephen Beveridge and Charles R Nelson. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary economics*, 7(2):151–174, 1981. 86

Natalia Z. Bielczyk, Sebo Uithol, Tim van Mourik, Paul Anderson, Jeffrey C. Glennon, and Jan K. Buitelaar. Disentangling causal webs in the brain using functional magnetic resonance imaging: A review of current approaches. *arXiv preprint*, 2017. 62

Juan A. Bonachela, Haye Hinrichsen, and Miguel A. Munoz. Entropy estimates of small data sets. *arXiv preprint*, 2008. doi: 10.1088/1751-8113/41/20/202001. 24

Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An Introduction to Transfer Entropy*. Springer-Verlag GmbH, 2016. ISBN 3319432214. 5, 13, 14, 23, 59

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLOS ONE*, 12(6):e0177678, 2017. doi: 10.1371/journal.pone.0177678. 66

Peter Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer New York, 2009. ISBN 1441903194. 15, 17

Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting (Springer Texts in Statistics)*. Springer, 2010. ISBN 0-387-95351-5. 15, 36, 43

Henk Broer and Floris Takens. *Dynamical Systems and Chaos*. Springer New York, 2011. doi: 10.1007/978-1-4419-6870-8. 52, 54

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 2020. doi: 10.1186/s12864-019-6413-7. 66

Adam Thomas Clark, Hao Ye, Forest Isbell, Ethan R. Deyle, Jane Cowles, G. David Tilman, and George Sugihara. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96(5):1174–1181, 2015. doi: 10.1890/14-1479.1. 62

Joshua N. Cooper and Christopher D. Edgar. A development of continuous-time transfer entropy. *arXiv preprint*, 2019. 9

Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006. ISBN 0471241954. 5, 6, 32

Fabian Dablander and Max Hinne. Node centrality measures are a poor substitute for causal inference. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-43033-9. 86

Graciela De Pierris and Michael Friedman. Kant and Hume on causality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2018. 11

Yu. G. Dmitriev and F. P. Tarasenko. On the estimation of functionals of the probability density and its derivatives. *Theory of Probability & Its Applications*, 18(3):628–633, 1974. doi: 10.1137/1118083. 25

J. L. Doob. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935. doi: 10.1214/aoms/1177732594. 24

Michael Eichler. Causal inference in time series analysis. In *Causality*, pages 327–354. John Wiley & Sons, Ltd, 2012. doi: 10.1002/9781119945710.ch22. 12

Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987. 43

Luca Faes, Giandomenico Nollo, and Alberto Porta. Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83(5), 2011. doi: 10.1103/physreve.83.051112. 67

Luca Faes, Silvia Erla, Alberto Porta, and Giandomenico Nollo. A framework for assessing frequency domain causality in physiological time series with instantaneous effects. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1997):20110618, 2013a. doi: 10.1098/rsta.2011.0618. xi, 74

Luca Faes, Giandomenico Nollo, and Alberto Porta. Compensated transfer entropy as a tool for reliably estimating information transfer in physiological time series. *Entropy*, 15(1):198–219, 2013b. doi: 10.3390/e15010198. 38

Andrea Falcon. Aristotle on causality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2019. 11

Conor Finn and Joseph T. Lizier. Generalised measures of multivariate information content. *Entropy*, 22(2):216, 2020. doi: 10.3390/e22020216. 15, 85

Ronald A Fisher. *The design of experiments*. Oliver & Boyd, 1949. 11

E. Fix and J.L. Hodges. Discriminatory analysis - nonparametric discrimination: consistency properties. *International Statistical Reviews*, 57:238–247, 1989. 20

Jack McKay Fletcher and Thomas Wennekers. From structure to activity: Using centrality measures to predict neuronal activity. *International Journal of Neural Systems*, 28(02):1750013, 2018. doi: 10.1142/s0129065717500137. 86

Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36, 2011. doi: 10.1089/brain.2011.0008. 59

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Estimating mutual information by local gaussian approximation. *arXiv preprint*, 2015. 26

Deniz Gencaga, Kevin Knuth, and William Rossow. A recipe for the estimation of information flow in a dynamical system. *Entropy*, 17(1):438–470, 2015. doi: 10.3390/e17010438. 9

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982. doi: 10.1080/01621459.1982.10477803. 13

John F. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984. doi: 10.1080/01621459.1984.10477110. 13

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x. 24

German Gomez-Herrero, Wei Wu, Kalle Rutanen, Miguel C. Soriano, Gordon Pipa, and Raul Vicente. Assessing coupling dynamics from an ensemble of time series. *arXiv preprint*, 2010. doi: 10.3390/e17041958. 31

S. V. Gonchenko, I. I. Ovsyannikov, C. Simó, and D. Turaev. Three-dimensional Hénon-like maps and wild Lorenz-like attractors. *International Journal of Bifurcation and Chaos*, 15(11): 3493–3508, 2005. doi: 10.1142/s0218127405014180. 54

S.V. Gonchenko, A.S. Gonchenko, I.I. Ovsyannikov, and D.V. Turaev. Examples of Lorenz-like attractors in Hénon-like maps. *Mathematical Modelling of Natural Phenomena*, 8(5):48–70, 2013. doi: 10.1051/mmnp/20138504. 54

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 72

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint*, 2017. 72

Carlos Granero-Belinchón, Stéphane G. Roux, and Nicolas B. Garnier. Information theory for non-stationary processes with stationary increments. *Entropy*, 21(12):1223, 2019. doi: 10.3390/e21121223. 31, 84

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424, 1969. doi: 10.2307/1912791. 12

Clive W. J Granger. Time series analysis, cointegration, and applications. *American Economic Review*, 94(3):421–425, 2004. doi: 10.1257/0002828041464669. 12

Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980. 12

Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373, 1988. doi: 10.1016/0375-9601(88)90193-4. 25, 27

Shuixia Guo, Anil K. Seth, Keith M. Kendrick, Cong Zhou, and Jianfeng Feng. Partial Granger causality—eliminating exogenous inputs and latent variables. *Journal of Neuroscience Methods*, 172(1):79–93, 2008. doi: 10.1016/j.jneumeth.2008.04.011. 72

Daniel Hahs and Shawn Pethel. Transfer entropy for coupled autoregressive processes. *Entropy*, 15(3):767–788, 2013. doi: 10.3390/e15030767. 35, 36, 37, 38

James Hamilton. *Time Series Analysis*. Princeton University Press, 1994. ISBN 0691042896. 13

Zengchao Hao and Vijay Singh. Integrating entropy and copula theories for hydrologic modeling and analysis. *Entropy*, 17(4):2253–2280, 2015. doi: 10.3390/e17042253. 86

Stefan Haufe, Vadim V. Nikulin, Klaus-Robert Müller, and Guido Nolte. A critical assessment of connectivity measures for EEG data: A simulation study. *NeuroImage*, 64:120–133, 2013. doi: 10.1016/j.neuroimage.2012.09.036. 73

Fei He, Stephen A. Billings, Hua-Liang Wei, and Ptolemaios G. Sarrigiannis. A nonlinear causality measure in the frequency domain: Nonlinear partial directed coherence with applications to EEG. *Journal of Neuroscience Methods*, 225:71–80, 2014. doi: 10.1016/j.jneumeth.2014.01.013. 73

Rainer Hegger, Holger Kantz, Lorenzo Matassini, and Thomas Schreiber. Coping with non-stationarity by overembedding. *Physical Review Letters*, 84(18):4092–4095, 2000. doi: 10.1103/physrevlett.84.4092. 86

Michel Hénon. A two-dimensional mapping with a strange attractor. In *The Theory of Chaotic Attractors*, pages 94–102. Springer, 1976. 53

Bradford Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 1965. 11

Katerina Hlavackova-Schindler, Milan Palus, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007. doi: 10.1016/j.physrep.2006.12.004. 13, 23

Meng Hu and Hualou Liang. A copula approach to assessing Granger causality. *NeuroImage*, 100: 125–134, 2014. doi: 10.1016/j.neuroimage.2014.06.013. 86

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018. 86

Ryan G. James, Nix Barnett, and James P. Crutchfield. Information flows? a critique of transfer entropies. *arXiv preprint*, 2015. doi: 10.1103/PhysRevLett.116.238701. 15

Ma Jian. Estimating transfer entropy via copula entropy. *arXiv preprint*, 2019. 86

Meihui Jiang, Xiangyun Gao, Haizhong An, Huajiao Li, and Bowen Sun. Reconstructing complex network for characterizing the time-varying causality evolution behavior of multivariate time series. *Scientific Reports*, 7(1), 2017. doi: 10.1038/s41598-017-10759-3. 86

Harry Joe. Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697, 1989. doi: 10.1007/bf00057735. 25

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202. 86

A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1-2):43–62, 2002. doi: 10.1016/s0167-2789(02)00432-3. 27, 35

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv preprint*, 2018. 72

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2006. ISBN 0521529026. 10, 29

Pavel Kindlmann and Francoise Burel. Connectivity measures: a review. *Landscape Ecology*, 2008. doi: 10.1007/s10980-008-9245-4. 73

Gebhard Kirchgässner and Jürgen Wolters. *Introduction to Modern Time Series Analysis*. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-73291-4. 18, 72

A.M. Kowalski, M.T. Martin, A. Plastino, and L. Zunino. Information flow during the quantum-classical transition. *Physics Letters A*, 374(17-18):1819–1826, 2010. doi: 10.1016/j.physleta.2010.02.037. 30

L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector, 1987. 26

Jakub Kořenek and Jaroslav Hlinka. Causal network discovery by iterative conditioning: comparison of algorithms. *arXiv preprint*, 2018. doi: 10.1063/1.5115267. 51, 59

Anna Krakovská, Jozef Jakubík, Martina Chvosteková, David Coufal, Nikola Jajcay, and Milan Paluš. Comparison of six methods for the detection of causality in a bivariate time series. *Physical Review E*, 97(4), 2018. doi: 10.1103/physreve.97.042207. 51, 59, 70

Alexander Kraskov, Harald Stoegbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 2003. doi: 10.1103/PhysRevE.69.066138. 27, 76, 84

Seung-Woo Ku, UnCheol Lee, Gyu-Jeong Noh, In-Gu Jun, and George A. Mashour. Preferential inhibition of frontal-to-parietal feedback connectivity is a neurophysiologic correlate of general anesthesia in surgical patients. *PLoS ONE*, 6(10):e25155, 2011. doi: 10.1371/journal.pone.0025155. 30

Dimitris Kugiumtzis. Direct coupling information measure from non-uniform embedding. *arXiv Preprint*, 2013. doi: 10.1103/PhysRevE.87.062918. 55, 62, 68, 76

Lee Lady. Calculus for the Intelligent Person, 2005. 98

Joseph Lizier and Mika Rubinov. Multivariate construction of effective computational networks from observational data. *arXiv Preprint*, 2012. 67

Joseph T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 2014. doi: 10.3389/frobt.2014.00011. 33, 49

Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 2008. doi: 10.1103/PhysRevE.77.026110. 15, 86

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi: 10.1175/1520-0469(1963)020⟨0130:dnf⟩2.0.co;2. 51

Edward N Lorenz. *The essence of chaos*. University of Washington press, 1995. 52

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag GmbH, 2005. ISBN 3540401725. 71

Jian Ma and Zengqi Sun. Mutual information is copula entropy. *arXiv preprint*, 2008. 86

James Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305. Citeseer, 1990. 60

B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. doi: 10.1016/0005-2795(75)90109-9. 66

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989. 24

George Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 1955. 24

A. Mokkadem. Estimation of the entropy and information of absolutely continuous random variables. *IEEE Transactions on Information Theory*, 35(1):193–196, 1989. 25

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318–2321, 1995. doi: 10.1103/physreve.52.2318. 27

Yonathan Murin, Jeremy Kim, Josef Parvizi, and Andrea Goldsmith. SozRank: A new approach for localizing the epileptic seizure onset zone. *PLOS Computational Biology*, 14(1):e1005953, 2018. doi: 10.1371/journal.pcbi.1005953. 86

Meinard Müller. Dynamic time warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74048-3_4. 86

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. doi: 10.3390/make1010019. xi, 59, 62, 72, 74

Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag GmbH, 2007. ISBN 0387286594. 86

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. *Statistical Science*, 5(4):465–472, 1923. ISSN 08834237. 11

J. M. Nichols, M. Seaver, S. T. Trickey, M. D. Todd, C. Olson, and L. Overbey. Detecting nonlinearity in structural systems using the transfer entropy. *Physical Review E*, 72(4), 2005. doi: 10.1103/physreve.72.046217. 35

Leonardo Novelli, Patricia Wollstadt, Pedro Mediano, Michael Wibral, and Joseph T. Lizier. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, 3(3):827–847, 2019. doi: 10.1162/netn_a_00092. 35

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272. 23, 24

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. Simulation study of direct causality measures in multivariate time series. *Entropy*, 15(12):2635–2661, 2013. doi: 10.3390/e15072635. 51, 59

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. Detecting causality in non-stationary time series using partial symbolic transfer entropy: Evidence in financial data. *Computational Economics*, 47(3):341–365, 2015. doi: 10.1007/s10614-015-9491-x. 11, 30, 31

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. ISBN 0-521-77362-8. 11, 14, 59, 63

Ernesto Pereda, Rodrigo Quian Quiroga, and Joydeep Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2):1–37, 2005. doi: 10.1016/j.pneurobio.2005.10.003. 73

David Martin Powers. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2011. 66

Stephen Ranshous, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F. Samatova. Anomaly detection in dynamic networks: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3):223–247, 2015. doi: 10.1002/wics.1347. 86

Riccardo Rossi, Andrea Murari, and Pasquale Gaudio. On the potential of time delay neural networks to detect indirect coupling between time series. *Entropy*, 22(5):584, 2020. doi: 10.3390/e22050584. 72

Kenneth Rothman. Causes. *American Journal of Epidemiology*, 104(6):587–592, 1976. doi: 10.1093/oxfordjournals.aje.a112335. 11

P. Rubenstein, S. Bongers, B. Scholkopf, and J. Mooij. From deterministic ODE's to dynamic structural causal models. *Uncertainty in Artificial Intelligence*, 2018. 52

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350. 11

J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018. doi: 10.1063/1.5025050. 69

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *arXiv preprint*, 2020. 69

Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical Review Letters*, 108(25), 2012. doi: 10.1103/physrevlett.108.258701. 67

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1), 2019a. doi: 10.1038/s41467-019-10105-3. 51, 59

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11), 2019b. doi: 10.1126/sciadv.aau4996. 62

Said E. Said and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984. doi: 10.1093/biomet/71.3.599. 58

V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, 41(12):1110–1117, 2011. doi: 10.1016/j.compbiomed.2011.06.020. 73

Steven J. Schiff, Paul So, Taeun Chang, Robert E. Burke, and Tim Sauer. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Physical Review E*, 54(6):6708–6724, 1996. doi: 10.1103/physreve.54.6708. 55

Katerina Schindlerova. Equivalence of granger causality and transfer entropy: A generalization. *Applied Mathematical Sciences, Hikari*, 2011. 14

Alois Schlögl. A comparison of multivariate autoregressive estimators. *Signal Processing*, 86(9): 2426–2429, 2006. doi: 10.1016/j.sigpro.2005.11.007. 74

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 2000. doi: 10.1103/PhysRevLett.85.461. 9

Alex Serès, Ana Alejandra Cabaña, and Argimiro Alejandro Arratia Quesada. Towards a sharp estimation of transfer entropy for identifying causality in financial time series. In *Proceedings of the 1st Workshop on MIning DAta for financial applicationS (MIDAS 2016), Riva del Garda, Italy, September 19-23, 2016*, pages 31–42, 2016. 14

Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 (3):379–423, 1948. 5

Elsa Siggiridou, Christos Koutlis, Alkiviadis Tsimpiris, and Dimitris Kugiumtzis. Evaluation of granger causality measures for constructing networks from multivariate time series. *Entropy*, 21 (11):1080, 2019. doi: 10.3390/e21111080. 51, 55, 59

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986. ISBN 0412246201. 19

Sara Sommariva, Alberto Sorrentino, Michele Piana, Vittorio Pizzella, and Laura Marzetti. A comparative study of the robustness of frequency–domain connectivity measures to finite data length. *arXiv preprint*, 2017. 73

Colin Sparrow. *The Lorenz Equations*. Springer New York, 1982. ISBN 0387907750. 51

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441, 1904. doi: 10.2307/1422689. 31

Richard E. Spinney, Mikhail Prokopenko, and Joseph T. Lizier. Transfer entropy in continuous time, with applications to jump and neural spiking processes. *Physical Review E*, 95(3), 2017. doi: 10.1103/physreve.95.032319. 9

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer New York, 2000. doi: 10.1007/978-1-4612-2748-9. 69

Olaf Sporns. *Networks of the Brain*. The MIT Press, 2010. ISBN 0262014696. 59

Matthäus Staniek and Klaus Lehnertz. Symbolic transfer entropy. *Physical Review Letters*, 100 (15), 2008. doi: 10.1103/physrevlett.100.158101. 30

Simon Streicher and Carl Sandrock. Plant-wide fault and disturbance screening using combined transfer entropy and eigenvector centrality analysis. *arXiv Preprint*, 2019. 86

G. Sugihara, R. May, H. Ye, C. h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012. doi: 10.1126/science.1227079. 70, 71

Fuqiang Sun, Wendi Zhang, Ning Wang, and Wei Zhang. A copula entropy approach to dependence measurement for multiple degradation processes. *Entropy*, 21(8):724, 2019. doi: 10.3390/e21080724. 86

Floris Takens. Detecting strange attractors in turbulence. In *Lecture Notes in Mathematics*, pages 366–381. Springer Berlin Heidelberg, 1981. doi: 10.1007/bfb0091924. 10, 70

J. B. Tenenbaum. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. 86

Y. L. Tong. *The Multivariate Normal Distribution*. Springer New York, 1990. doi: 10.1007/978-1-4613-9655-0. 101

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/cbo9780511802256. 20

Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1):45–67, 2010. doi: 10.1007/s10827-010-0262-3. 11

Ioannis Vlachos and Dimitris Kugiumtzis. Nonuniform state-space reconstruction and coupling detection. *Physical Review E*, 82(1), 2010. doi: 10.1103/physreve.82.016207. 60, 68

Vincent Q. Vu, Bin Yu, and Robert E. Kass. Information in the non-stationary case. *Neural Computation*, 2008. 29

Huifang E. Wang, Christian G. Bacnar, Pascale P. Quilichini, Karl J. Friston, Viktor K. Jirsa, and Christophe Bernard. A systematic framework for functional connectivity measures. *Frontiers in Neuroscience*, 8, 2014. doi: 10.3389/fnins.2014.00405. 73

Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T. Lizier, and Raul Vicente. Measuring information-transfer delays. *PLoS ONE*, 8(2):e55809, 2013. doi: 10.1371/journal.pone.0055809. 31

N. Wiener. The theory of prediction. Modern Mathematics for Engineers, 1956. 12

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint*, 2010. 15, 67

Patricia Wollstadt, Mario Martínez-Zarzuela, Raul Vicente, Francisco J. Díaz-Pernas, and Michael Wibral. Efficient transfer entropy analysis of non-stationary neural time series. *PLoS ONE*, 9 (7):e102833, 2014. doi: 10.1371/journal.pone.0102833. 31

Patricia Wollstadt, Joseph Lizier, Raul Vicente, Conor Finn, Mario Martinez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. IDTxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34):1081, 2019. doi: 10.21105/joss.01081. 62, 67

Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921. 11

Hao Ye, Ethan R. Deyle, Luis J. Gilarranz, and George Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5(1), 2015. doi: 10.1038/srep14750. xi, 70, 74

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint*, 2015. 72

Appendix A

Theoretical supplements

The theoretical results that were omitted from the main text are listed here.

The choice of logarithm in Shannon entropy

The entropy of a discrete random variable is a real number. Therefore it makes sense to interpret entropy as an expectation of a function of the random variable, say $E[g(X)]$ where g is a suitably chosen function. The Law of the Unconscious Statistician yields that $E[g(X)] = \sum_x p_X(x)g(x)$. It will turn out to be convenient to use a slightly less general form through replacing $g(x)$ by $g(p_X(x))$ (there is a slight loss of generality here since p_X need not be an invertible function). We will use a similar functional form for the joint entropy. We thus consider:

$$\begin{aligned} H(X) &= \sum_x p_X(x)g(p_X(x)) \\ H(Y) &= \sum_y p_Y(y)g(p_Y(y)) \\ H(X,Y) &= \sum_x \sum_y p_{X,Y}(x,y)g(p_{X,Y}(x,y)) \end{aligned}$$

We further require that the joint entropy of two independent random variables is the sum of the entropies of the individual random variables:

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

Then, this requirement implies the additivity of function g:

$$\begin{aligned} \sum_x \sum_y p_{X,Y}(x,y)g(p_X(x)\cdot p_Y(y)) &= \sum_x 1\cdot p_X(x)g(p_X(x)) + \sum_y 1\cdot p_Y(y)g(p_Y(y)) \\ &= \sum_x \sum_y p_Y(y)p_X(x)g(p_X(x)) + \sum_y \sum_x p_X(x)p_Y(y)g(p_Y(y)) \\ &= \sum_x \sum_y p_{X,Y}(x,y)\left(g(p_X(x)) + g(p_Y(y))\right) \end{aligned}$$

Comparing the first and the last sum we get:

$$g(p_X(x)\cdot p_Y(y)) = g(p_X(x)) + g(p_Y(y))$$

This additive property characterizes the function g within the class of measurable functions, and a straightforward proof of that can be given provided we make the extra assumption that g is differentiable and defined for positive real numbers. The following argument is based on Lee Lady (2005), while for the general proof we refer to Aczél and Daróczy (1975). Assuming the above conditions for g we have:

If $g(x \cdot y) = g(x) + g(y)$, then $g(1) = g(1 \cdot 1) = g(1) + g(1)$, therefore $g(1) = 0$. Fixing an arbitrary $x > 0$ we then compute the derivative:

$$g'(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \tag{A.1}$$

$$= \lim_{h \to 0} \frac{g(x) + g(1 + \frac{h}{x}) - g(x)}{h} \tag{A.2}$$

$$= \lim_{h \to 0} \frac{g(1 + \frac{h}{x}) - g(1)}{h} \tag{A.3}$$

$$= \frac{1}{x} \lim_{u \to 0} \frac{g(1+u) - g(1)}{u} \tag{A.4}$$

$$= \frac{1}{x} g'(1) \tag{A.5}$$

where in (A.4) we set $u = h/x$. So, for an arbitrary $x > 0$ we have:

$$g(x) = g'(1) \log x + c$$

Re-using $g(1) = 0$ we obtain $c = 0$. Writing $p_X(x)$ instead of x we conclude:

$$g(p_X(x)) = k \log p_X(x)$$

where $k = g'(1)$. Since $p_X(x) \leq 1$ and we would want the information function g (and, equivalently, the entropy H) to be non-negative, $k = g'(1)$ should be chosen negative. This also implies that the derivative $g'(p_X(x)) < 0$, i.e. that g is decreasing: less probable outcomes yield more information. In the main definition given in Chapter 2, $k = -1$. This choice serves a normalization purpose, as in that case, the entropy of the outcome of the toss of a fair coin is 1 bit.

Proof of Theorem 2.1.10

The proof of the non-negativity of the K-L divergence is included here.

Proof. Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Note that because of the condition assumed for p and q in Definition 2.1.9, $p(x) > 0$ implies $q(x) > 0$, i.e. the following are well defined. We write:

$$-D(p\|q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{A.6}$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} = \log \sum_{x \in A} q(x) \tag{A.7}$$

$$\leq \log \sum_{x \in \mathcal{X}} q(x) = \log 1 = 0 \tag{A.8}$$

(A.7) follows from Jensen's inequality for concave functions: $\phi\left(\sum a_i x_i\right) \geq \sum a_i \phi(x_i)$, where $\phi(x) = \log(x)$, $a_i = p(x)$ and $x_i = q(x)/p(x)$.

Since $\log t$ is a strictly concave function of t, we have equality in (A.7) iff $q(x)/p(x)$ is constant everywhere, i.e. $q(x) = cp(x)$. Therefore, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$.

Equality in (A.8) holds only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p\|q) = 0$ iff $p(x) = q(x)$ for all x. □

Proof of Theorem 2.1.18

The proof of the analytic expression for transfer entropy is included here.

Proof. Recall the expression of TE as a difference of conditional entropies (2.21):

$$T_{Y\to X}^{(k,\ell)}(t) = H(X_t|X_{t-1}^{(k)}) - H(X_t|X_{t-1}^{(k)}, Y_{t-1}^{(\ell)})$$

Then from the definition of conditional entropy (2.6), we have:

$$H(X_t|X_{t-1}^{(k)}) = -\sum_{x_{t-1}^{(k)}} p(x_{t-1}^{(k)}) \sum_{x_t} p(x_t|x_{t-1}^{(k)}) \log p(x_t|x_{t-1}^{(k)})$$

$$H(X_t|X_{t-1}^{(k)}, Y_{t-1}^{(\ell)}) = -\sum_{x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} p(x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \sum_{x_t} p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)})$$

We now use the definition of conditional probability $p(x|y) = p(x,y)/p(y)$ for $p(x_t|x_{t-1}^{(k)})$ and $p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)})$. Then, we note that the denominator of the resulting fractions is independent of x_t, and therefore it can be taken out of the inner sum and cancel with $p(x_{t-1}^{(k)})$ and $p(x_{t-1}^{(k)}, y_{t-1}^{(\ell)})$ respectively. Thus:

$$H(X_t|X_{t-1}^{(k)}) = -\sum_{x_{t-1}^{(k)}} \sum_{x_t} p(x_t, x_{t-1}^{(k)}) \log p(x_t|x_{t-1}^{(k)}) \tag{A.9}$$

$$= -\sum_{x_t, x_{t-1}^{(k)}} p(x_t, x_{t-1}^{(k)}) \log p(x_t|x_{t-1}^{(k)}) \tag{A.10}$$

$$\begin{aligned} H(X_t|X_{t-1}^{(k)}, Y_{t-1}^{(\ell)}) &= -\sum_{x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} \sum_{x_t} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \\ &= -\sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \end{aligned}$$

The result is now given by noting:

$$p(x_t, x_{t-1}^{(k)}) = \sum_{y_{t-1}^{(\ell)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)})$$

Therefore (A.10) becomes:

$$H(X_t|X_{t-1}^{(k)}) = -\sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)})$$

and the result follows:

$$\begin{aligned} T_{Y\to X}^{(k,\ell)}(t) &= -\sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)}) + \sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \log p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)}) \\ &= \sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(l)}} p(x_t, x_{t-1}^{(k)}, y_{t-1}^{(l)}) \log \frac{p(x_t|x_{t-1}^{(k)}, y_{t-1}^{(\ell)})}{p(x_t|x_{t-1}^{(k)})} \end{aligned}$$

□

Proof of Theorem 2.3.4

Here we prove that a strictly stationary time series with finite variance is weakly stationary.

Proof. Setting $k = 1$ in (2.42) yields, for any $t \in \mathbb{Z}$ and any $h \in \mathbb{Z}$

$$X_t \overset{d}{=} X_{t+h}$$

For an arbitrary $s \in \mathbb{Z}$, setting $h = s - t$, the above proposition proves that for any $t, s \in \mathbb{Z}$ we have $X_t \overset{d}{=} X_s$. Since equality in distribution implies equality of means, we get $E[X_t] = m$ for all $t \in \mathbb{Z}$.

Setting $k = 2$ in (2.42) yields, for any $t, s \in \mathbb{Z}$ and any $h \in \mathbb{Z}$

$$(X_t, X_s) \overset{d}{=} (X_{t+h}, X_{s+h}) \tag{A.11}$$

Due to $E\big[|X_t|^2\big] < \infty$ for all $t \in \mathbb{Z}$, the Cauchy-Schwarz inequality implies that all covariances are finite: $\big(Cov(X_t, X_s)\big)^2 \leq Var(X_t)Var(X_s) < +\infty$. Then, (A.11) implies

$$\begin{aligned} Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h}) \iff \\ \gamma_X(t, s) = \gamma_X(t + h, s + h) \end{aligned}$$

Therefore, X_t is weakly stationary. □

An example of a weakly but not strictly stationary time series is subsequently given here.

Example A.0.1. *Consider a sequence of independent random variables $\{X_t, t \in \mathbb{Z}\}$ such that for t odd the distribution of X_t is exponential with mean equal to one, while for t even the distribution of X_t is normal with mean and variance equal to one:*

$$X_t \sim \begin{cases} Exp(1) & , \text{ if } t \text{ is odd} \\ \mathcal{N}(1, 1) & , \text{ if } t \text{ is even} \end{cases}$$

Then, $E[X_t] = 1$ and $E[X_t^2] = 2 < +\infty$ for any $t \in \mathbb{Z}$. Also, for any $t, s \in \mathbb{Z}$

$$\gamma_X(t, s) = Cov(X_t, X_s) = E[X_t X_s] - E[X_t]E[X_s] = E[X_t X_s] - 1$$

For an arbitrary $h \in \mathbb{Z}$:

$$\text{If } t = s, \text{ then } Cov(X_t, X_s) = Var(X_t) = 1 = Var(X_{t+h}) = Cov(X_{t+h}, X_{s+h})$$

Otherwise:

- *if h is even, $X_{t+h} \overset{d}{=} X_t$ and $X_{s+h} \overset{d}{=} X_s$. Hence, $E[X_{t+h}X_{s+h}] \overset{indep.}{=} E[X_{t+h}]E[X_{s+h}] = E[X_t]E[X_s] \overset{indep.}{=} E[X_t X_s]$*
- *If h is odd, $X_{t+h} \overset{d}{=} X_s$ and $X_{s+h} \overset{d}{=} X_t$. Hence, $E[X_{t+h}X_{s+h}] \overset{indep.}{=} E[X_{t+h}]E[X_{s+h}] = E[X_s]E[X_s] \overset{indep.}{=} E[X_s X_t]$*

In any case, $E[X_t X_s] = E[X_{t+h}X_{s+h}]$. This implies

$$Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h})$$

That is, X_t is weakly stationary. However, since e.g. X_1 and X_2 have different distributions, X_t is not strictly stationary.

Multivariate normal distribution theory

We then consider the definition and characteristic function of a multivariate normal random variable as well as relevant properties.

Definition A.0.2. *A random variable X of dimension n follows a multivariate normal distribution with mean μ and (non-singular) covariance matrix $\Sigma = \Sigma_{X,X}$, written as $X \sim N(\mu, \Sigma)$ if it has the following probability density function:*

$$f_X(x) = (2\pi)^{-\frac{n}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\big[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\big] \tag{A.12}$$

Theorem A.0.3. *The characteristic function of a multivariate normal random variable $X \sim N(\mu, \Sigma)$ admits the following form:*

$$\phi_X(t) = E[e^{it^T X}] = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right) \tag{A.13}$$

Since characteristic functions characterize the probability distribution of the random variable involved and (A.13) depends only on the mean μ and covariance matrix Σ, we infer that a multivariate normal distribution is fully specified by its mean and covariance matrix.

An important property of multivariate normality is the fact that it is preserved under affine transformations. For the proof of the following proposition we refer to (Tong, 1990, p. 32).

Theorem A.0.4. *If $X \sim N(\mu, \Sigma)$ is an n-dimensional normal random variable, B is a $m \times n$ real matrix and a is a $m \times 1$ real vector, then the random vector $Y = a + BX$ is again multivariate normal. Specifically:*

$$Y \sim N(a + B\mu, B\Sigma B^T) \tag{A.14}$$

As a corollary, any subset of a multivariate normal vector X is also multivariate normal distributed:

Corollary A.0.5. *Let $X \sim N(\mu, \Sigma)$ an n-dimensional normal random variable $X = (X_1, ..., X_n)$. For any set of indices $i_1, ..., i_k$, $k < n$ the multivariate random variable $X' = (X_{i_1}, ..., X_{i_k})$ is normally distributed.*

Proof of Theorem 4.1.6

Proof that, as defined in Theorem 4.1.6, time series V_t is white noise.

Proof. Referring to Definition 4.1.3 we have to prove the following:

- V_t is stationary
- The mean of V_t is zero: $\mu_t = 0$
- The covariance matrix of V_t maps non-zero arguments to zero: $\Gamma(h) = 0$ if $h \neq 0$

The processes Z_t, W_t are both i.i.d noise with zero mean therefore it is trivially seen that $E[V_t] = 0$ (and it also does not depend on t, a condition needed for stationarity). We then compute the covariance matrix of V_t as defined in (4.1.2):

$$\Gamma(t+h, t) = \begin{bmatrix} Cov(Z_{t+h}, Z_t) & Cov(Z_{t+h}, bZ_t + W_t) \\ Cov(bZ_{t+h} + W_{t+h}, Z_t) & Cov(bZ_{t+h} + W_{t+h}, bZ_t + W_t) \end{bmatrix} \tag{A.15}$$

Referring to the independence assumptions of the model, we note that, for $h \neq 0$, all covariances refer to different and therefore independent time points. Thus, when $h \neq 0$ all elements of the above matrix are 0 (and trivially independent from t). By calculating the specific covariances when $h = 0$ we therefore proved that V_t is a white noise process with covariance matrix $\Sigma = \Gamma(0)$. □

Appendix B

Additional graphs

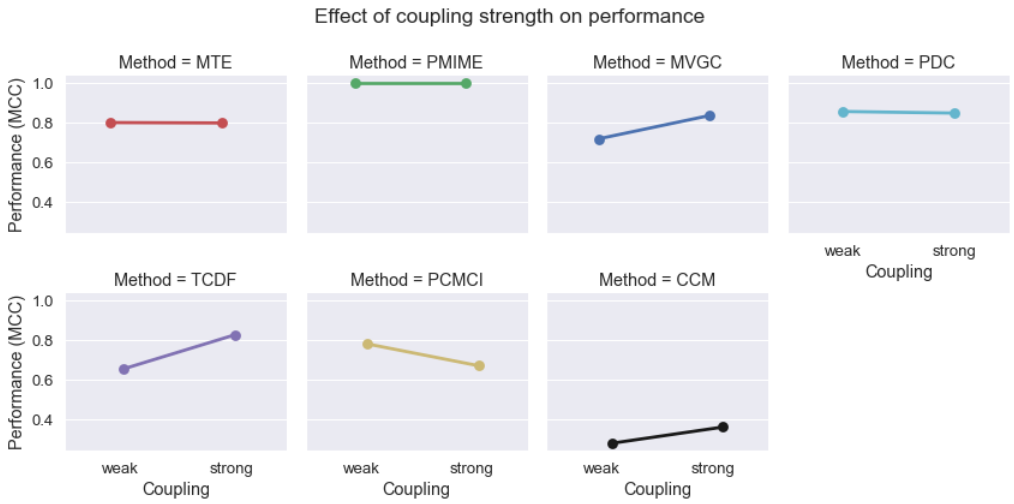

Figure B.1: The average performance of each method on data groups H_3 and H_1.

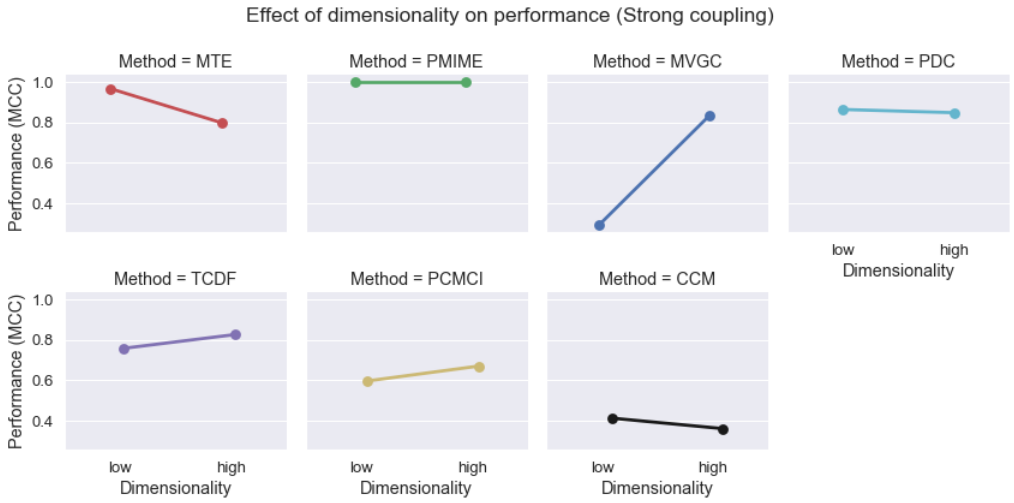

Figure B.2: The average performance of each method on data groups H_1 and H_2.

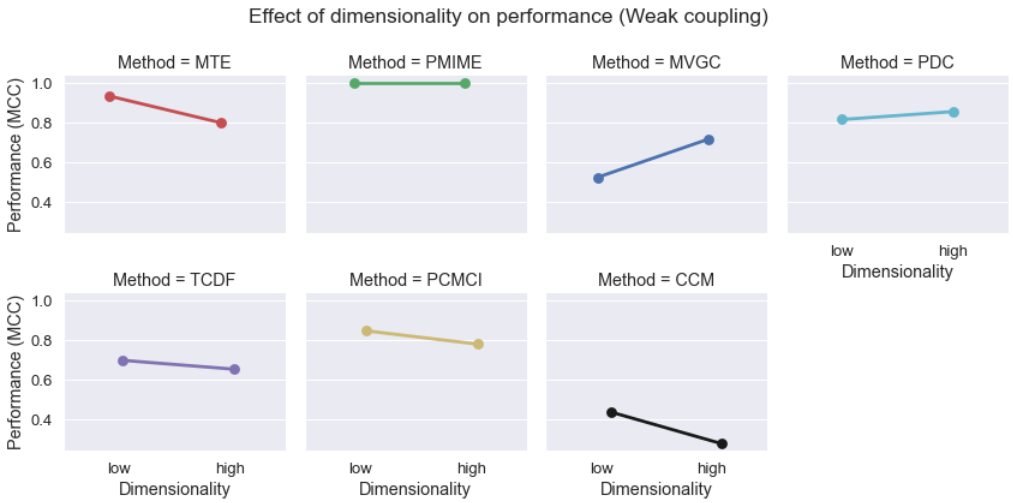

Figure B.3: The average performance of each method on data groups H_3 and H_4.

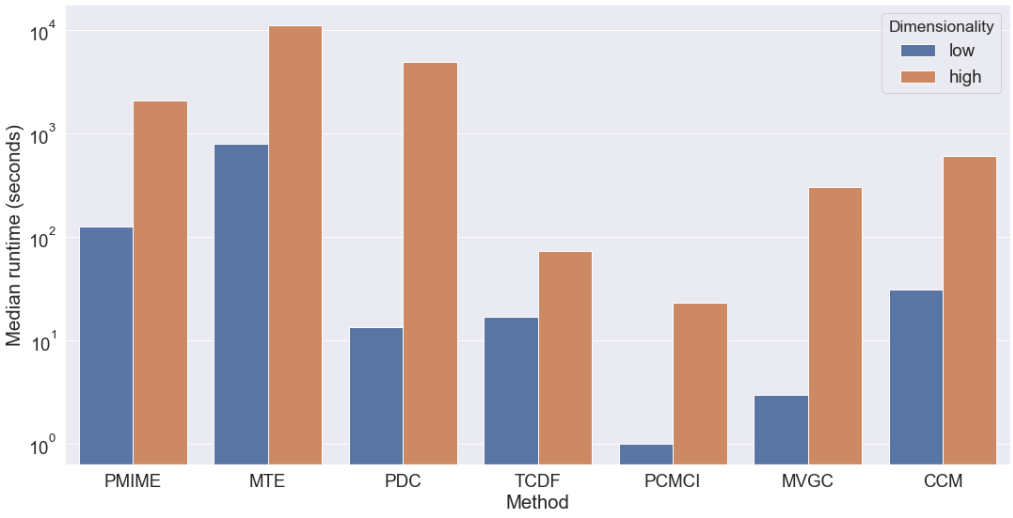

Figure B.4: The difference in average median runtime between low and high dimensional datasets for each method (log-scale).