# A Novel Causal Inference Method for Time Series

MASTER THESIS OF: NAJI SHAJARISALES

ADVISORS: MICHEL BESSERVE
DOMINIK JANZING

MAX-PLANCK-GESELLSCHAFT

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

# ACKNOWLEDGEMENTS

I greatly appreciate Michel Besserve's helps during my master thesis. He was not only a supervisor but a caring friend that helped me to pave my way through for the whole duration. I also thank Dominik Janzing for supervising me and all the fruitful discussions that I had with him. I thank my parents Fariba and Mohammad and my sister Sana, for their 24/7 support and I thank all my friends who listened to my complaints without complaining back.

# CONTENTS

# LIST OF FIGURES

# 1 INTRODUCTION

## 1.1 MOTIVATING PROBLEMS

Suppose that we have a dataset of observed pair of measurements $x_i, y_i$ for $0 \leqslant i \leqslant n$ from two observables X and Y. Moreover we are guaranteed that the these observables are causally related in the sense that either X causes Y, which we represent with $X \to Y$ or Y causes X which we represent with $Y \to X$. The task is to differentiate the cause from effect.

To clarify what we mean by observables we give a few different examples. Suppose we are given $n$ different pairs of texts such that elements of each pair are composed of an original text in English (German) and a translation by Google translate to German (English). A human fluent in both languages can easily identify the cause (the real text) from the effect (the translated text) in all the cases without being hindered, even if $n = 1$. But yet, there is no possible computational solution that could do the same task as good as a human agent.

A more numerical example would be $n$ pairs of recorded temperature and height measurements in different places in a geographic area. Its known a priori that change of height influences the temperature but not vice versa. However we are interested in a computational solution that could answer the same question using observed data without an a priori knowledge of their origin explicitly given to the computational solver.

Finally we explain an example from neuroscience. Recordings of brain electrical activity at multiple scales, including Local Field Potentials, Electrocorticograms and Electroencephalograms are important techniques for analyzing and understanding the underlying brain activity [14] . One of the chief applications of these methods is the localization of the focal region, responsible for triggering seizures in patients with medically refractory focal epilepsy [38]. An application of causal inference is to find the directional interactions between different areas of the brain from the recorded signals, and infer the source of pathological brain activity. This example is a problem of causal inference for time series, which usually calls for other specific algorithm than causal inference for individual random variables.

A novel framework of causal inference has been recently established in our laboratory [33], relying on the postulate of independence of cause and mechanism (ICM). However, so far, no inference method dedicated to time series has been developed based on this framework. The objective of our

work is to propose theoretical foundations and algorithms for such a method. Since this will enable us to study how the methods based on the ICM postulate compares to other causal inference methods in the context of time series.

## 1.2 OUTLINE

In the following, the first chapter introduces the necessary mathematical terminology and material.

The second chapter will give an overview on the history of causality and causal inference. It also gives a brief introduction to available causal inference methods in machine learning. Then it focuses on an overview of causal inference methods for time series problems (when observables as described in previous section are changing over time). Finally we give a brief overview of causality methods for time series that are applied to neural data and discuss some of their shortcomings.

In the third chapter we introduce a new framework which is based on the following intuitive assumption that has been elaborated there:

"*the cause is in some way independent from the mechanism that generates the effect from cause.*"

In chapter four we formulate the above principle in a mathematical framework for these observables that are *discrete* time series.

In chapter five we derive some identifiability results for this method; meaning that we present assumptions that based on them one can derive cause from effect on theoretical grounds. Moreover we sketch some connections of this method to already established methods based on the intuitive independence assumption stated above, namely causal inference method for observables related through high-dimensional linear relationships and causal inference method for observables that are related through nonlinear relationships.

In chapter six we focus on the case where time series are Gaussian processes and relate the method to a proposed causal inference scheme known as Information Geometrical Causal Inference (IGCI) [35].

In chapter seven we apply our method to synthesized data and to real world data and measure its effectiveness in practice.

Chapter eight has been dedicated to discussion over the results and about the future works on this new causal inference method.

We have decided to keep part of the results developed in appendix; In this part we explain another way of deriving the inference method proposed in this thesis, with a different approach and using the toolkits of linear operator theory.

# 2

# BASIC TERMINOLOGY AND PRELIMINARIES

This chapter introduces the necessary mathematical toolkit for the upcoming chapters. For the sake of clarity, we chose to gather here the necessary basic notions and results from different subjects and state them in their own section. Moreover in this way the reader will always have the chance to refer to this section to find the necessary information, for later-to-be-seen uses of these terms and preliminary mathematical toolkits.

## 2.1 SOME NOTATIONS

We represent any $N$-dimensional complex (real) vector with bold characters as $\mathbf{z} := (z_1, ..., z_N)$. For $\mathbf{x} \in \mathbb{C}^N$, $\|\mathbf{x}\|_p$ represents the vector norm in $\mathbb{C}^N$ which is defined as:

$$\|\mathbf{x}\|_p = \begin{cases} \sqrt[p]{\sum_{i=0}^{N-1} |x_i|^{1/p}} & \text{if } p \in (0, +\infty) \\ \max_{0 \leqslant i \leqslant N-1} |x_i| \end{cases}$$

where $|x|$ represents the modulus (absolute value) of $x$. Also $\bar{\mathbf{x}}$ (or equivalently $\mathbf{x}^*$) represents the complex conjugate of $\mathbf{x}$. For any given pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in vector spaces $V_1$ and $V_2$, $\mathbf{x}{:}\mathbf{y} \in V_1 \oplus V_2$ is the vector derived from concatenating $\mathbf{x}$ and $\mathbf{y}$. For a given $k \in \mathbb{R}$ and interval $I \subset \mathbb{R}$, $L^k(I)$ represents the set of measurable functions $f$ such that the Lebesgue integral $\int_I |f(x)|^k dx$ is finite. We represent the space of real (complex) valued $m \times n$ matrices with $M_{m\times n}(\mathbb{R})$ (and equivalently $M_{m\times n}(\mathbb{C})$). For a given matrix $A$, we represent the element in row $i$ and column $j$ with $[A]_{ij}$.

## 2.2 PROBABILITY THEORY

We introduce here, some preliminaries from probability theory by means of measure theory. For more on this one can refer to [39]. We believe this gives a more rigour picture of all the probabilistic arguments in this thesis.

**Definition 1.** *[39]*

*(i) A pair $(\Omega, \mathcal{A})$ consisting of a nonempty set $\Omega$ and a $\sigma$-algebra $\mathcal{A} \subset 2^\Omega$ is called a **measurable space**. The sets $A \in \mathcal{A}$ are called **measurable sets**. If $\Omega$ is at most countably infinite and if $\mathcal{A} = 2^\Omega$ , then the measurable space $(\Omega, 2^\Omega)$ is called **discrete**.*

*(ii) A triple $(\Omega, \mathcal{A}, \mu)$ is called a **measure space** if $(\Omega, \mathcal{A})$ is a measurable space and if $\mu$ is a measure on $\mathcal{A}$.*

*(iii) If in addition $\mu(\Omega) = 1$, then $(\Omega, \mathcal{A}, \mu)$ is called a **probability space**.*

**Definition 2. (Random variables)**. *[39] Let $(\Omega, \mathcal{A})$ be a measurable space and let $X : \Omega \to \Omega'$ be measurable. $X$ is called a **random variable** with values in $(\Omega', \mathcal{A}')$. If $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where $\mathcal{B}(\mathbb{R})$ is the Borel algebra of $\mathbb{R}$ then $X$ is called a real random variable or simply a random variable. Complex random variables are defined in the same way.*

For a given random variable $X$ from $(\Omega, \mathcal{A})$ with measure $P$ to $(\Omega', \mathcal{A}')$, $X_*P$ is the induced measure over $(\Omega', \mathcal{A}')$ defined as:

$$\forall A' \in \mathcal{A}, \quad X_*P(A') = P(X^{-1}(A'))$$

Unless otherwise stated all the random variables have the same probability space as domain $(\Omega, \mathcal{A}, P)$.

**Definition 3. (Absolute Continuity for Measures)** *For two given measures $\mu$ and $\nu$ over measurable space $(\Omega, \mathcal{A})$, we say $\mu$ is absolutely continuous with respect to $\nu$ or $\nu \ll \mu$ if*

$$\nu(A) = 0 \quad \textit{for all } A \in \mathcal{A} \ \textit{ with } \ \mu(A) = 0.$$

A measure over $(\Omega, \mathcal{A})$ is called $\sigma$-finite if there exist a sequence $\Omega_i \in \mathcal{A}$ such that $\bigcup \Omega_i = \Omega$ and $\mu(\Omega_i) < \infty$ for any $i$.

**Theorem 1. (Radon-Nykodim Derivative)** *Let $\mu$ and $\nu$ be $\sigma$-finite measures on $(\Omega, \mathcal{A})$ and moreover suppose $\nu \ll \mu$. Then there exist a function $f : \Omega \to [0, \infty)$ called the Radon-Nikodyim derivative of $\mu$ with respect to $\nu$ that satisfies*

$$\forall A \in \mathcal{A}, \quad \mu(A) = \int_A d\mu = \int_A f d\nu.$$

*One represents $f$ usually with $\frac{d\mu}{d\nu}$. $f$ is also called the density of $\mu$ with respect to $\nu$.*

Suppose $X$ is a random variable with domain measure $P$ and values in $(\Omega', \mathcal{A}')$ and given the reference measure $\mu$ over $\mathcal{A}'$. If $\mu \ll X_*P$ then density with respect to $\mu$ is defined as the Radon-Nikodym derivative $\frac{dX_*P}{d\mu}$. When $\mu$ is the Lebesgue measure, we write $\frac{dX_*P}{d\mu}$ as $p_X(x)$ and this is the so called probability density function. Throughout the document for a given random variable $X$, $P_X$ represents its measure over domain and $p_X$ represents its density with respect to Lebesgue measure in case of existence.

For a given (complex) random variable $X$, we define its expected value denoted by $\mathbb{E}_P(X)$ as

$$\mathbb{E}_P(X) := \int_\Omega X dP.$$

For two complex random variables $X$ and $Y$ and a given reference measure $\mu$, the standard covariance function is defined [29, p. 376] as

$$\mathrm{Cov}_\mu(X,Y) := \int_\Omega X\bar{Y}d\mu - \int_\Omega X d\mu \int_\Omega \bar{Y} d\mu$$

As a special case the variance of a complex valued random variable $X$ with respect to measure $\mu$ is defined as

$$\mathrm{Var}_\mu(X) := \mathrm{Cov}_\mu(X,X).$$

In all the cases unless there is no confusion on the reference measure we omit the subscript of the measure in the above definitions.

## 2.3 TIME SERIES

### 2.3.1 *Deterministic Time Series*

We refer to a sequence of real numbers $\{x_t, t \in T\}$ (which can also be sequence of complex numbers and even a sequence of vectors) as a deterministic time series. When the index set $T$ is already known we use the shorter notation of $\{x_t\}$. We interchangeably use sequence notations to refer to deterministic time series, or equivalently we might also consider $x(t)$ as a function of $t$ which in some cases represents the time.

For a deterministic time series, when $x(t)$ is a function on $\mathbb{R}$, i.e. $T = \mathbb{R}$, if $x(t) \in L^1(\mathbb{R})$, we represent the Fourier transform [25, pp. 155] of it denoted as $\hat{x}_\nu$ (or $\mathcal{F}(\{x_t\})$) with:

$$\hat{x}_\nu = \int_{-\infty}^{\infty} x_t e^{-2\pi i \nu t} dt,$$

and in case $\hat{x}(\nu)$ is in $L^1(\mathbb{R})$ then $x$ is called its inverse Fourier transform and one has [25, pp. 155]

$$x_t = \int_{-\infty}^{\infty} \hat{x}_\nu e^{2\pi i \nu t} d\nu.$$

When $T = \mathbb{Z}$ (a.k.a discrete time series) we use the following definition for Fourier transform with the same notation if $\{x_t\} \in l^1(\mathbb{Z})$:

$$\forall -\frac{1}{2} \leqslant \nu \leqslant \frac{1}{2} \quad \hat{x}_\nu = \sum_{-\infty}^{\infty} x_t e^{-2\pi i \nu t},$$

and when $\hat{x}_\nu \in L^1([-\frac{1}{2}, \frac{1}{2})$ then the inverse Fourier transform is defined as:

$$\forall t \in \mathbb{N} \quad x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{x}_\nu e^{-2\pi i \nu t} d\nu.$$

Plancherel theorem states that when $\{x_t\}$ and $\{y_t\}$ are in $L^2(\mathbb{R})$ ($T = \mathbb{R}$), then

$$\int_{-\infty}^{\infty} x(t)y(t)dt = \int_{-\infty}^{\infty} \hat{x}(\nu)\hat{y}(\nu)d\nu,$$

which also holds for discrete case when the integral on both sides are replaced with proper sums. For a given deterministic time series, $C_x(\tau)$ will represent its autocorrelation function:

$$C_x(\tau) := \int_{-\infty}^{\infty} x(t)x(t+\tau)dt.$$

And we define the energy spectral density to be:

$$S_{xx}(\nu) := |\hat{x}(\nu)|^2 \tag{1}$$

Then one can show that

$$C_X(\tau) = \int_{-\infty}^{\infty} e^{2\pi i\tau\nu} S_{xx}(\nu)d\nu$$
$$S_{xx}(\nu) = \int_{-\infty}^{\infty} e^{-2\pi i\tau\nu} C_x(\tau)d\tau.$$

We use $\Psi(\{x_t\})$ to represent the overall energy of $x(t)$

$$\Psi(\{x_t\}) = \int_{-\infty}^{\infty} |x_t|^2 \stackrel{*}{=} \int_{-\infty}^{\infty} |\hat{x}_\nu|^2 = \int_{-\infty}^{\infty} S_{xx}(\nu)d\nu = C_x(0)$$

where the (*) is based on Plancherel's theorem.

For any two deterministic time series $\{x_t\}$ and $\{y_t\}$ the convolution between $\{x_t\}$ and $\{y_t\}$ denoted as $\{x_t\} * \{y_t\}$ is a function on the same index set $T$ defined as

$$(\{x_t\} * \{y_t\})(\tau) = \int_{-\infty}^{\infty} x_\tau y_{t-\tau} dt$$

### 2.3.2 *Stochastic Time Series and Processes*

Most of the terminology with regard to time series and stochastic processes are from [12]. A *stochastic process* is a family of random variables $\{X_t, t \in T\}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ and $T$ is the index set. For any random variable $X : \Omega \to E$ and $\omega \in \Omega$, $X(\omega)$ is one realization of $X$. Similarly for a given $\omega \in \Omega$ the family $\{X_t(\omega)\}$ is a realisation of the stochastic process. For a given index $t$ we use $\{X_t\}$ to represent the random variable at index $t$. In what follows except when explicitly stated otherwise, $T = \mathbb{Z}$. We use $\{X_t\}$ to represent the complete stochastic process and we use $X_{t:s}$ to indicate the time series between two time instances when there exist a nat-

ural order over index set. We also use $\{\tilde{X}_t\}$ to represent the reverse process, i.e. $\{\tilde{X}_t\}$ is the process where

$$\forall t \in \mathbb{Z} \quad \tilde{X}_t = X_{-t}.$$

We use the same notation for time reversal of a sequence.

Throughout we may assume that stochastic processes considered are all zero mean, i.e. for any $\{X_t\}$, $\mathbb{E}_P(\{X_t\}) = 0$. In cases where our index set $T$ is $\mathbb{Z}$, we only consider stochastic time series that are **purely non-deterministic** unless explicitly stated otherwise. A purely nondeterministic process is defined as follows:

**Definition 4.** *[32, p. 88]* **(Purely nondeterministic processes)** *For a given stochastic process $\{X_t\}$, take $\mathcal{H}_n(X)$ to be the subspace of $L^2(\Omega, P)$ spanned by $X_k$ for $k \leqslant n$. A weakly stationary process is said to be deterministic if*

$$\mathcal{H}_n(X) = \mathcal{H}_{n+1}(X), \quad n \in \mathbb{Z}$$

*and purely nondeterministic if*

$$\bigcap_{n\in\mathbb{Z}} \mathcal{H}_{n+1}(X) = \{0\}, \quad n \in \mathbb{Z}.$$

**Remark 1.** *[32] It can be shown that a weakly stationary process $X$ is purely non-deterministic if and only if the spectral distribution function is absolutely (see theorem 2 for definition) continuous with respect to Lebesgue measure and its SDF (see theorem 2 for definition) $S_{xx}(\nu)$ satisfies*

$$\left| \int_{-\frac{1}{2}}^{\frac{1}{2}} \log(S_{xx}(\nu)) d\nu \right| < \infty.$$

For a given stochastic process $\{X_t\}$ and given time instances $t, s$ the autocovariance function $C_X(t, s)$ is defined as

$$C_X(t, s) = \mathbb{C}ov_P(X_t, X_s) = \mathbb{E}_P(X_t X_S) - \mathbb{E}_P(X_t)\mathbb{E}_P(X_S) = \mathbb{E}_P(X_t X_s),$$

based on the assumption that stochastic processes considered are zero mean. Similarly, the mean of a time series $\mu_X$ is defined as:

$$\forall t, \quad \mu_X(t) := \mathbb{E}_P(X_t).$$

Its important to note that none of these functions are necessarily defined for all time instances. A process is called zero mean if

$$\forall t, \quad \mu_X(t) = 0.$$

In the context of time series analysis a specific family of stochastic processes known as *stationary processes* plays a fundamental role. It is defined as follows:

**Definition 5. (Weak stationarity)***[12] The time series $\{X_t, t \in \mathbb{Z}\}$, is said to be weakly stationary (or stationary in wide sense) if*

*(i)* $\mathbb{E}_P|X_t|^2 < \infty$ *for all* $t \in \mathbb{Z}$,

*(ii)* $\mu_X(t) = m$ *for all* $t \in \mathbb{Z}$

*(iii)* $C_X(r, s) = C_X(r + t, s + t)$ *for all* $s, t, r \in \mathbb{Z}$

**Remark 2.** *For a weakly stationary time series since the autocovariance function is invariant under shift of time (condition (iii) in definition 5) we represent the autocovariance function with a single argument:*

$$C_X(\tau) := \mathbb{E}_P\left[X_t X_{t+\tau}\right]$$

*where $\tau$ is said to be the lag of the autocovariance function.*

For a given deterministic time series $\{x_t\}$, the Fourier transform provides a representation of time series known as frequency domain representation. For the purpose of this thesis we want to generalize this frequency domain representation to stochastic processes; for weakly stationary stochastic processes when $T = \mathbb{Z}$ (and even when $T = \mathbb{R}$) one has the following theorem known as Wiener-Khintchine theorem.

**Theorem 2. (Wiener-Khintchine theorem)***[17, pp. 95] Suppose a real-valued weakly stationary process $\{X_t\}$ is given where $C_X(\tau)$ exists for every lag $\tau$. Then there exist a monotonically increasing function $F$ defined on $[-\frac{1}{2}, \frac{1}{2})$ such that*

$$C_X(\tau) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \tau \nu} dF(\nu).$$

*Moreover since $\{X_t\}$ is purely nondeterministic, $F$ is absolutely continuous with respect to the Lebesgue measure. Take $S_{xx}(\nu)$ to denote this density. Then*

$$C_X(\tau) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \tau \nu} S_{xx}(\nu) d\nu$$

$$S_{xx}(\nu) = \sum_{k=-\infty}^{\infty} e^{-2\pi i k \nu} C_X(\tau)$$

$S_{xx}(\nu)$ is the density function associated with $F$ which is called **spectral density function (SDF)**. This means that $C_X$ and $S_{xx}$ are Fourier transform pairs. For a zero mean weakly stationary stochastic time series $\{X_t\}$ we define $P(X_t)$ to be the power of the time series as

$$P(X_t) = C_X(0) = \mathbb{E}_P(X_t^2) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu) d\nu$$

**Remark 3.** *We might as well use other equivalent terms for this SDF such as power spectral density or simply power spectrum interchangeably.*

We will also need the notion of **white noise process**. A white noise process is a weakly stationary process where the autocovariance function is zero for any nonzero argument. A Gaussian white noise process is a white noise process where the underlying finite joint distributions are Gaussian.

### 2.3.3 *The special case of finite circular time series*

For the purpose of this thesis, it is interesting to draw a link between infinite weakly stationary time series and finite random sequences with similar invariance properties. We thus introduce the necessary formalism to define circular translation invariant sequences. For this we elaborate on spectral properties of deterministic and weakly stationary finite time series.

A special case of Fourier transform known as Discrete Fourier Transform (DFT) for a finite time series $\{x_t\}$ where $0 \leqslant t \leqslant N-1$ is defined as

$$\forall 0 \leqslant k \leqslant N-1 \quad \hat{x}_k = \sum_{k=0}^{N-1} x_n e^{-2\pi i k \frac{n}{N}}.$$

It can be shown that there exist a unitary matrix $U_{\mathcal{F}}$ such that for any time series of length $N$ represented as a vector $\mathbf{x} = \{x_t\}$, one has

$$\hat{\mathbf{x}} = \sqrt{N} U \mathbf{x},$$

where $\hat{\mathbf{x}}$ is the vector representation of $\{\hat{x}_\nu\}$. For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, one can define the **circular convolution** of these two $\mathbf{z} = \mathbf{x} *_c \mathbf{y}$ as

$$z_t = \sum_{i=0}^{N-1} x_{(i)_N} y_{(t-i)_N}$$

where $(m)_N$ means $m$ modulo $N$. A circulant matrix $C \in M_{N\times N}(\mathbb{R})$ is a matrix where

$$\forall k \in \mathbb{Z} \quad [C]_{ij} = [C]_{(i-k)_N (j-k)_N]}$$

Fourier transform matrix $U_{\mathcal{F}}$ is circular and therefore it can be easily shown that $U_{\mathcal{F}}\mathbf{x}$ is equivalent to circular convolution of $\mathbf{x}$ with the first row of $U_{\mathcal{F}}$. It can be shown that $U_{\mathcal{F}}$ diagonalizes any circulant matrix, i.e. $U_{\mathcal{F}} C U_{\mathcal{F}}^\top$ is a diagonal matrix for any circulant matrix $C$. The diagonal elements in this case will be the eigenvalues of $C$ and as a result one has

$$\operatorname{tr}(U_{\mathcal{F}} C U_{\mathcal{F}}^\top) = \operatorname{tr}(C) \tag{2}$$

Based on this version of Fourier transform we derive an equivalent of Wiener-Khintchine theorem for stochastic processes with finite length. To

this end, we define a process $\{X_t\}$ which satisfies $X_t = X_{t+N}$ for a given $N$ and for any $t$. To impose such a structure on a process pick a random vector $\mathbf{Z}$ of dimension $N$ such that

$$\forall t \in \mathbb{Z}, \quad X_t = Z_{(t)_N}$$

In such a case $\{X_t\}$ is called a **circular process** [37]. Since all the information of this process is confined to $X_{0:N-1}$ we only consider this part of the process. We define Fourier transform for $X_{0:N-1}$ as $U_{\mathcal{F}}\mathbf{X}$ where $\mathbf{X}$ is $\mathbf{Z}$ with some overload of notation. If we name the covariance matrix associated with $\mathbf{X}$ as $\Sigma_X$, one can notice that the first row of this matrix exactly represents $C_X(k)$ (again for the same confined length). Furthermore for $C_X(k)$ seen as a vector it can be shown [26, pp. 202] that

$$\operatorname{diag}(U_{\mathcal{F}} C_X) = U_{\mathcal{F}} \Sigma_X U_{\mathcal{F}}^{\top},$$

where $\operatorname{diag}(v)$ is a diagonal matrix with elements of vector $v$ over its principal diagonal. Therefore representing the Fourier transform of $C_X$ with $S_{xx}$ we realize that the diagonal elements of $U_{\mathcal{F}} \Sigma_X U_{\mathcal{F}}^{\top}$ are nothing but $S_{xx}$. We consider $S_{xx}$ as the power spectral density of $\{X_t\}$. Moreover using eq. (2) we get

$$\frac{1}{n} \sum_{\nu=0}^{n-1} S_{xx}(\nu) = C_X(0).$$

### 2.3.4 *Gaussian Processes*

Finally since we investigate some properties only for special processes known as Gaussian Processes, we introduce them here. We restrict our attention to one dimensional real Gaussian processes, i.e. the processes where at each index, the random variable is real valued. In the following definition $T \in \{\mathbb{Z}, \mathbb{R}\}$.

**Definition 6.** *A stochastic process $\{X_t, t \in T\}$, is a Gaussian processes if the joint distribution associated with any finite set of indices is a multivariate Gaussian distribution. We define $\mathcal{K}_{xx} : T \times T \to \mathbb{R}$ as the covariance kernel (covariance function) of the process, i.e. the gram matrix of this kernel over $t_1, \ldots, t_n \in T$ is the covariance matrix of the multidimensional Gaussian associated with these indices in the process. Moreover $\mu_X : T \to \mathbb{R}$ is the mean function associated with the process, i.e.*

$$\mu_X(t) = \mathbb{E}_P(X(t)) = 0.$$

## 2.4 INFORMATION THEORY AND INFORMATION GEOMETRY

Measures of information theory have been proved to be of utmost importance in theoretical and applied sciences. One of the most important usage

of these measures is for quantifying independence between events and random variables. Since independence plays a crucial role in inferring causal relationships, information theory is an important tool in causal inference methods.

**Definition 7. (Relative Entropy)** *For two given probability measures* $P$ *and* $Q$ *over the same measure space if* $Q \ll P$ *the Kullback-Leibler, relative entropy, or KL divergence of* $P$ *with respect to* $Q$ *is defined as*

$$D_{KL}(P\|Q) = \int \log(\frac{dP}{dQ})dP$$

*where* $\frac{dP}{dQ}$ *is the Radon-Nikodym derivative.*

When both $P$ and $Q$ are absolutely continuous with respect to a reference measure $\mu$ (mostly Lebesgue) with densities $p_\mu$ and $q_\mu$ respectively and moreover $Q \ll P$, then we get:

$$D_{KL}(p\mu\|q_\mu) := D_{KL}(P\|Q) = \int \log(\frac{dP}{dQ})dP = \int p_\mu(x)\log(\frac{p_\mu(x)}{q_\mu(x)})d\mu(x)$$

And in case of the Lebesgue measure we get:

$$D_{KL}(p\|q) := \int p(x)\log(\frac{p(x)}{q(x)})dx$$

One can notice that the above definition includes the discrete case, where $\Omega$ is at most countably infinite space; then the integral turns to summation and density will be taken with respect to counting measure. This is true for any other integration throughout the article. Based on this definition we introduce a notion from information geometry namely orthogonality in information space.

**Definition 8. (Orthogonality in Information Space)** *For given densities* $p(x), q(x), r(x)$ *defined over a given space* $\mathcal{X}$*, we say* $(p, q, r)$ *makes a Pythagorean triple iff*

$$D_{KL}(p\|q) = D_{KL}(p\|r) + D_{KL}(r\|q).$$

*Then it is said that the vector connecting* $p$ *to* $r$ *is orthogonal to the vector connecting* $r$ *to* $q$*, when densities are seen as infinite dimensional vectors.*

Finally we state a lemma that will be later used to extend the observation of theorem 7 to discrete Gaussian processes. Before, we need to define the notion of relative entropy rate.

**Definition 9.** *[32] Let* $X = \{X_t\}$ *and* $Y = \{Y_t\}$ *be discrete stochastic processes. The* ***relative entropy rate*** $\bar{D}(P_{X_t}\|P_{Y_t})$ *is defined as*

$$\bar{D}(P_{X_t}\|P_{Y_t}) := \lim_{N\to\infty} \frac{1}{2N+1} D_{KL}(P_{X_{-N:N}}\|P_{Y_{-N:N}})$$

*where* $P_{X_{-N:N}}$ *stands for the joint measure over* $X_{-N}, \ldots X_N$.

**Lemma 1.** *[32] Let* $X = \{X_n : n \in \mathbb{Z}\}$ *and* $Y = \{Y_n : n \in \mathbb{Z}\}$ *be zero mean purely nondeterministic weakly stationary discrete Gaussian processes with SDF's* $S_{xx}$ *and* $S_{yy}$*, respectively. Then the relative entropy rate is given by*

$$\bar{D}(P_{X_n} \| P_{Y_n}) = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \Big( \frac{S_{xx}(\nu)}{S_{yy}(\nu)} - 1 - \log \frac{S_{xx}(\nu)}{S_{yy}(\nu)} \Big) d\nu$$

*provided that at least one of the following conditions is satisfied.*

*(i)* $\frac{S_{xx}(\nu)}{S_{yy}(\nu)}$ *is bounded*

*(ii)* $S_{yy}(\nu) > a > 0$ *for all* $\nu \in [-\pi, \pi]$ *and* $S_{xx} \in L^2[-\pi, \pi]$

## 2.5 LINEAR SYSTEMS

For any input output system, a main concern is to model the system behaviour, i.e. to be able to predict the output of the system given a history of input [and maybe outputs]. We will consider the case of **deterministic** system; this means that there exist a function $\mathcal{S}$ associated with the system such that:

$$\mathcal{S}(\{x_t\}) = \{y_t\}$$

where $\{x_t\}$ and $\{y_t\}$ are real valued inputs and outputs of the system respectively. More specifically we mean that each value $\{y_t\}$ is a function of all the $\{x_t\}$ values; in a context where $t$ represents time, this indicates a system (function) that depends on past and/or future of the input. When the output of a filter depends only on the past inputs to the filter the filter is called a **causal** filter(system). **Linear systems** are a special case of this family which are extensively studied partly for the reason that their study is analytically more tractable and partly because they can provide a framework for the study of nonlinear systems [17, 62]. A system is called linear if for any given pairs of time series $\{x_t^1\}$ and $\{x_t^2\}$ as inputs and for any real $\alpha$ and $\beta$ we have:

$$\mathcal{S}(\alpha\{x_t^1\} + \beta\{x_t^2\}) = \alpha\mathcal{S}(\{x_t^1\}) + \beta\mathcal{S}(\{x_t^2\})$$

Moreover we focus our attention over systems whose behaviour is invariant under time shift known as *time(shift)-invariant systems*. More precisely a system is said to be time invariant if, assuming $\{x_t\}$ and $\{y_t\}$ are input and output pairs of $\mathcal{S}$ for all $t$

$$\forall \tau \in \mathbb{Z}, \quad \mathcal{S}(\{x_{t-\tau}\}) = \{y_{t-\tau}\}.$$

The systems satisfying both properties are known as Linear Time Invariant(LTI) systems (filters). The LTI filters with discrete time inputs and out-

puts are the main subject of our study in section 4.2.1.

**Proposition 1.** *[46] For any LTI system there exist a function known as* ***impulse response function*** $h(t)$ *such that for any* $\{x_t\}$ *where one has* $\{y_t\} = \mathcal{S}(\{x_t\})$ *then*

$$\forall \tau \in \mathsf{T} \;\; y_\tau = (\{x_t\} * \{h_t\})(\tau).$$

We may denote the impulse response function with $h_t$ or simply as $h$ interchangeably. When the space of inputs two an LTI filter is based on a weakly stationary process then one can also derive the following:

**Proposition 2.** *[46] For a weakly stationary stochastic process* $\{X_t\}$ *the output of an LTI system* $\mathcal{S}$ $\{Y_t\}$ *is weakly stationary as well. Moreover if* $\Psi(\{h_t\}) < \infty$ *then*

$$\forall \nu \in \mathbb{R} \quad S_{yy}(\nu) = S_{xx}(\nu)|\hat{h}(\nu)|^2$$

Proposition 2 is closely connected with some of the statements in appendix A.1 since autocovariance functions for weakly stationary processes are positive definite functions.

**Remark 4.** *We confined our discussion to systems with real valued inputs and outputs mostly for the sake of clarity. Most of the properties indicated here can have equivalents for inputs and outputs belonging to other domains, e.g. vectors and complex number. Same is true for the index set that has been taken to be integer numbers but it can be e.g. continuous like* $\mathbb{R}$ *and then under some mild conditions the same properties will hold again.*

We also define the notion of inverse filter; for an LTI $\mathcal{S}$, when the transfer function (Fourier transform of impulse response function) is not vanishing, one can show that the filter that takes as its input the output of $\mathcal{S}$ and generates its input which we denote with $\mathcal{S}^{-1}$, has an impulse response function $h^{(-1)}$ such that

$$\widehat{h^{(-1)}}_\nu = \frac{1}{\hat{h}_\nu}.$$

We call such a filter the **inverse filter** associated with $\mathcal{S}$.

### 2.5.1 *IIR and FIR systems*

Since some of the results in this thesis are based on an special type of filters known as Infinite Impulse Response (IIR) and Finite Impulse Response (FIR) filters, we introduce these type of filters explicitly. When the impulse response function $h_t$ of an LTI is only nonzero in finitely many points then the filter is known as FIR filter. The notion of IIR filters is closely related to

digital filters. Digital filters that are heavily used in signal processing discipline are usually defined in terms of difference equations, i.e. the equations of the form:

$$y[n] = \frac{1}{a_0}\left(\sum_{i=0}^{P} b_i x[n-i] + \sum_{j=1}^{Q} a_j y[n-j]\right).$$

One can check that a filter based on this definition is LTI. Such LTI filters are known as Infinite Impulse Response systems (filters). For these filters P is known as feedforward order which we shortly represent as FO, Q is feedback order which we abbreviate to BO. $a_i$'s and $b_i$'s are known as feedback and feedforward coefficients respectively. We will interchangeably represent these coefficients as vectors $\mathbf{a}$ and $\mathbf{b}$ and as a pair $(\mathbf{a}, \mathbf{b})$. When all the $a_i$'s (except $a_0$) are zero the filter is a causal FIR filter. In the context of stochastic time series, if the input to the IIR filter ($X_i$'s) would be white noise then such a data generating model is called an Auto Regressive Moving Average (ARMA) model.

## 2.6 LINEAR ALGEBRA

We confine our discussions here to linear operators that are known as integral operators and also to linear operators known as infinite dimensional matrices. We start by defining these two terms.

**Definition 10. (Integral Operator)** *[7, pp.58] Let $\mathcal{T}_{\mathcal{L}} : C^0[a,b] \to C^0[a,b]$ be a linear map, where $\mathcal{L} : [a,b] \times [a,b] \to \mathbb{R}$ is a continuous function. For a given $f \in C[a,b]$, define $g = \mathcal{T}_{\mathcal{L}}(f)$ as*

$$\forall y \in [a,b], \quad g(y) = \mathcal{T}_{\mathcal{L}}(f)(y) := \int_a^b \mathcal{L}(y,x) f(x)\,dx \tag{3}$$

*where $\mathcal{L}$ is called the kernel of $\mathcal{T}_{\mathcal{L}}$.*

With some overload of notation, in case there is no confusion we refer to $\mathcal{L}$ as the linear operator and kernel itself. In a quite similar way one can define the infinite dimensional matrices.

**Definition 11. (Infinite Dimensional Matrix)** *Let $\mathcal{T}_{\mathcal{L}} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ be a linear map and also assume $\mathcal{L} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ is a function. For a given deterministic time series $\{x_t, t \in \mathbb{Z}$, $\mathcal{T}_{\mathcal{L}}(\{x_t\})$ is defined as*

$$\forall \tau \in \mathbb{Z}, \quad y_\tau = \mathcal{T}_{\mathcal{L}}(\{x_t\}) = \sum_{k=-\infty}^{\infty} \mathcal{L}(\tau, k) x_k, \tag{4}$$

*where $\mathcal{L}$ is called an infinite dimensional matrix.*

Again with some overload of notation we use $\mathcal{L}(\{x_t\})$ instead of $\mathcal{T}_{\mathcal{L}}(\{x_t\})$.

We also can define the composition of two linear operators say $\mathcal{L}$ and $\mathcal{K}$ as

$$\forall x, y \in \mathsf{T} \quad (\mathcal{K}\mathcal{L})(x, y) := \int \mathcal{K}(x, z)\mathcal{L}(z, y)dz$$

where $\mathsf{T}$ is either $\mathbb{Z}$ or $\mathbb{R}$ in our case, and in the former the integral becomes summation. It can be seen as in the case of matrices, $\mathcal{L}$ acts on $\mathcal{K}$ from the right and therefore on its second argument while on the other hand $\mathcal{K}$ acts on $\mathcal{L}$ from left and therefore it acts on the first argument of $\mathcal{L}$ in the integral (summation in discrete case). In this sense its similar to the difference of act of linear operator over an input from the left and right. There too, $(f)\mathcal{L}$ and $(\{x_t\})\mathcal{L}$ are defined in the same way as eqs. (3) and (4) except that integration and summation takes place over the second argument respectively.

A linear operator $\mathcal{L}$ is said to be **shift invariant** if $\mathcal{L}(t, s) = \mathcal{L}(t + \tau, s + \tau)$ for all $t$, $s$ and $\tau$ in $\mathbb{Z}$. For a shift invariant operator $\mathcal{L}$ we define a function $\Phi_{\mathcal{L}}$ as

$$\forall t \in \mathbb{Z} \quad \Phi_{\mathcal{L}}(t) = \mathcal{L}(0, t). \tag{5}$$

This function carries all the information about $\mathcal{L}$, i.e. one can also construct $\mathcal{L}$ from $\Phi_{\mathcal{L}}$ with eq. (5), since $\mathcal{L}$ is shift invariant. For a linear operator $\mathcal{L}$, $\mathcal{L}^\top$ is the unique operator satisfying

$$\langle \mathcal{L}f, g\rangle = \langle f, \mathcal{L}^\top g\rangle$$

known as transpose of $\mathcal{L}$. In the next two lemmas we show that the space of shift invariant operators is closed under transposition and under composition of shift invariant operators.

**Lemma 2. (Closedness under transpose)** *For a shift invariant linear operator $\mathcal{L}$, $\mathcal{L}^\top$ is shift invariant and moreover one has*

$$\Phi_{\mathcal{L}^\top} = \tilde{\Phi}_{\mathcal{L}}.$$

*Proof.* Take any $f \in l^2(\mathbb{Z})$ and $g \in l^2(\mathbb{Z})$. We have

$$\langle \mathcal{L}f, g\rangle = \langle \Phi_{\mathcal{L}} * f, g\rangle = \sum_{t=-\infty}^{\infty} g(t) \sum_{\tau=-\infty}^{\infty} \Phi_{\mathcal{L}}(t-\tau)f(\tau) = \sum_{t=-\infty}^{\infty} g(t) \sum_{\tau=-\infty}^{\infty} \tilde{\Phi}_{\mathcal{L}}(\tau-t)f(\tau)$$

and changing the order of summation gives

$$\langle \mathcal{L}f, g\rangle = \sum_{\tau=-\infty}^{\infty} f(\tau) \sum_{t=-\infty}^{\infty} \tilde{\Phi}_{\mathcal{L}}(\tau-t)g(t) = \langle f, \tilde{\Phi}_{\mathcal{L}} * g\rangle \tag{6}$$

Now define $\mathcal{L}'$ to be the operator associated with $\tilde{\Phi}_{\mathcal{L}}$. Uniqueness of transpose of an operator and Equation (6) shows that $\mathcal{L}' = \mathcal{L}^{\top}$. □

**Lemma 3. (Closedness under composition)** *For any pair of translation invariant linear operators $\mathcal{L}$ and $\mathcal{H}$, $\mathcal{L}\mathcal{H}$ is translation invariant and we have*

$$\Phi_{\mathcal{L}\mathcal{H}} = \Phi_{\mathcal{L}} * \Phi_{\mathcal{H}}$$

*Proof.* Take any $f$ in the domain of $\mathcal{H}$. Then

$$\forall y \in \mathbb{Z} \quad \mathcal{L}\mathcal{H}(f)(y) = \mathcal{L}\left(\sum_{x=-\infty}^{\infty} \mathcal{H}(\cdot, x) f(x)\right)(y) =$$

$$\sum_{z=-\infty}^{\infty} \mathcal{L}(y, z) \sum_{x=-\infty}^{\infty} \mathcal{H}(z, x) f(x) = \sum_{z=-\infty}^{\infty} \Phi_{\mathcal{L}}(y - z)(\Phi_{\mathcal{H}} * f)(z) =$$

$$(\Phi_{\mathcal{L}} * \Phi_{\mathcal{H}} * f)(y)$$

where the final equation follows from associativity of convolution operation. □

**Definition 12.** *For any infinite dimensional matrix with kernel $\mathcal{L}$, we define its normalized trace $\mathcal{L}$ as*

$$\mathcal{T}_{\mathcal{L}} = \lim_{N \to \infty} \frac{\sum_{k=-N}^{N} \mathcal{L}(k, k)}{2N + 1}$$

*when this limits exists. For an integral operator $\mathcal{L}$ over $\mathbb{R}$ define the normalized trace as*

$$\mathcal{T}_{\mathcal{L}} = \lim_{\tau \to \infty} \frac{\int_{-\tau}^{\tau} \mathcal{L}(t, t) dt}{2\tau}$$

*when the above limit exists.*

**Remark 5.** *For a linear operator with translation invariant kernel $\mathcal{L}$, normalized trace is always defined and its equal to $\Phi_{\mathcal{L}}(0)$.*

A real valued function $\phi$ is positive definite if the shift invariant Kernel defined as $K(x, y) = \phi(x - y)$ is positive definite.

**Theorem 3.** ***(Bochner's Theorem).*** *[56] A continuous complex-valued function $\phi$ on $\mathbb{R}$ is positive definite if and only if it can be represented as*

$$\phi(\tau) = \int_{\mathbb{R}} e^{2\pi i \nu \tau} d\mu_{\phi}(\nu) \tag{7}$$

*where $\mu$ is a positive finite measure.*

**Corollary 1.** *Suppose the assumptions of Bochner's theorem hold and suppose that $\mu_\phi$ has a density $S_\phi$ with respect to the Lebesgue measure. Then*

$$\phi(0) = \int_{\mathbb{R}} e^{2\pi i \nu \tau} d\mu_\phi(\nu) = \int S_\phi(\nu) d\nu$$

$S_\phi$ is called the spectral density of $\phi$.

**Remark 6.** *All of the results above can be extended to continuous linear operators with change of summation notation to integration when necessary.*

We will also need some results regarding the largest eigenvalues of matrices and the relation to the largest eigenvalues of their submatrices.

**Theorem 4.** *[61] Let $M$ be an $n \times n$ Hermitian matrix and $\lambda_1, ..., \lambda_n$ its eigenvalues arranged in increasing order, counted with multiplicity. Then*

$$\lambda_k = \min_{\dim F = k} \max_{x \in F \setminus \{0\}} \frac{\mathbf{x}^* M \mathbf{x}}{\|\mathbf{x}\|_2^2}$$

*where the minimum is taken over $k$-dimensional vector spaces of $\mathbb{C}$.*

Define $\rho(A)$ as the largest eigenvlue of a given hermitian matrix $A$. A submatrix of a matrix $A$, is a matrix obtained by deleting some rows and some columns of $A$. A submatrix is called principal if the deleted rows and columns have the same index. Then this follows from the above theorem.

**Corollary 2.** *For a given Hermitian matrix $H$ and any principal submatrix of $H$, $H'$ one has*

$$\rho(H) \geqslant \rho(H').$$

# 3

# SOME PRINCIPLES OF CAUSAL INFERENCE

In this chapter we give an overview on the history of causal inference and put emphasis on frameworks in the machine learning community. We then explain the limits of causal graphical models relying on the Markov condition [45, 67] and specifically causal faithfulness to derive causal relations and introduce a recent framework [33] that can address these problems. We explain two different approaches based on this framework. Finally we discuss the application of causal inference methods for time series and more specifically for neural data and point out some of their shortcomings.

## 3.1 HISTORY OF CAUSALITY AND CAUSAL INFERENCE

Causal inference is nowadays considered as one of the new branches of artificial intelligence [10] that has attracted a lot of attention from different disciplines of science in the last two decades, since causal relations can provide predictability under manipulation which is not the case for approaches based on statistical dependency, e.g. correlation. In the following section, we try to sketch a short introduction and history on causality and causal inference. A detailed introduction of causality can be found in [73]. For a chronological list of contributors to the field, one can refer to [2].

The early studies on causality date back to Aristotle and Plato, and it has since then been a topic of interest for many philosophers such as Kant, Leibniz, Hume, Russel to name only a few. It remains an important topic for philosophy hitherto which has developed into many branches [73]. As a first step towards experimental sciences, the systematic study of causality as a problem of inference based on observational data goes back to the seminal work of Sewall Wright [75], where he defines for the first time causal network models. This approach has been later developed mainly through the works of H. Simon [63–65], J. Pearl [45] and P. Spirtes [67]. The work of the latter two mainly relies on representing causal relations based on Bayesian nets. Pearl introduced the notion of Functional Causal Models and together with Spirtes advocated the application of two postulates: the *Causal Markov Condition* and *Causal Faithfulness* (with a different terminology in Pearl) into the field of causality. Its worthwhile to note that the application of Markov condition in causal inference is due to the seminal work of Hans Reichenbach relevant to the common cause principle [54, pp. 157].

Causal Markov condition states that for any Bayesian network with variable set $V$, its joint distribution $P$ satisfies Markov condition, i.e. the conditional distribution (based on $P$) of any node given its parents is independent

of its non-descendants in the Bayesian net [67, pp. 11]. Faithfulness which complements the Markov condition asserts that all the conditional independences of P are entailed by Markov condition [67, pp. 13]. Although frameworks relying on these two assumptions improved the science of causal discovery astonishingly, they also had their shortcomings.

One major problem is that applying these methods on considerably large Bayesian networks will usually infer large family of equivalent causal graphs that all are able to explain the observed independences without the possibility of further progress on identifying the ground truth causal graph. Even when the family of these equivalent classes is not large it can still be problematic; A striking example of this type is deciding the cause and effect in bivariate networks, i.e. $X \to Y$ or $Y \to X$.

The second problem is that these methods rely on the possibility of sampling in an identically and independently distributed (i.i.d) manner. Indeed, many real world problems have time-varying observations and cannot be well modelled using the i.i.d. assumption on observed data, or involve sample sizes so small that assessing statistical independence is challenging or impossible. A noticeable example of the latter issue is the causal inference based on single observations. Although the first problem can be addressed through available methods in classical statistics of inference on time dependant data, the second problem remains unsolved using these conventional methods. In the next section we introduce a framework that can naturally address these problems.

## 3.2 INDEPENDENCE OF CAUSE AND MECHANISM (ICM)

Based on a preliminary idea by Lemeire and Dirkx [42], Janzing and Schölkopf [33] established a new framework for causal inference based on the minimum description length principle. We illustrate an intuitive view of their theory: They assume that the cause and mechanism are independent in the sense that they have been chosen by nature through two different processes. But more precisely suppose we represent the random variable for cause with $C$ and the random variable for effect with $E$. They argue that among factorization of the joint probability of cause $C$ and effect $E$ $P(C, E)$, into $P(C|E)$ and $P(E)$, the factorization reflecting the underlying causal structure typically leads into simpler expressions of $P(C|E)$ and $P(E)$ in terms of minimum description length (Kolmogorov complexity) [16, 40, 66]; this on the other hand means that $P(E)$ and $P(C|E)$ in the correct causal model are algorithmically independent in terms of Kolmogorov complexity. Throughout this manuscript, we will refer to this framework as Independence of Cause and Mechanism (ICM). Since the Kolmogorov complexity is not computable, practical methods relying on the ICM postulate must resort to other, computable, complexity measures.

One of the main applications of ICM addresses the case of deterministic relation between the cause and effect, i.e. there exist a deterministic function $f$ such that $E = f(C)$. More precisely assuming that for two observed random variables $X$ and $Y$ where $Y = f(X)$ and the ground truth is either of $X \to Y$ or $Y \to X$, the objective is to identify the correct underlying causal relationship. Since in this case there is no noise, most of the available techniques to address causal inference for bivariate data like [30, 47, 76] become ineffective. Based on ICM, [20, 34, 35, 77] introduced methods for inferring causal directions in this scenario.

More specifically [34, 77] address the case where $f$ is a linear high dimensional function, i.e. $E = AC$ where $A$ is a matrix and $E$ and $C$ are multidimensional random variables. In this case the method exploits the covariance structure of the cause and effect vectors. [20] on the other hand assumes deterministic relations where $f$ is nonlinear and exploits the possible dependency between the non-linearity of $f$ and the distribution of cause and effect using information geometric measures. Finally [35] hints to a connection between these two ICM-based methods – for linear high-dimensional relations and nonlinear relations – through information geometry for the Gaussian case. We will give a brief overview of these methods in the next section.

### 3.2.1 *ICM for Deterministic Nonlinear Relations*

Assume a given pair of observed random variables $X$ and $Y$ with $Y = f(X)$ where $f$ is a nonlinear diffeomorphism and the objective is to find the ground truth which is either of $X \to Y$ and $Y \to X$.

To this end we also introduce $u_X$ and $u_Y$ as appropriate reference measures for $X$ and $Y$ defined on domains of $X$ and $Y$ (as functions) respectively. By reference measure here what we mean is a measure that is used to quantify the irregularities that are present in densities of the input and output observables and also the irregularities of the functional relationship between the two. And by appropriate reference measure we refer to measures that are based on the problem domain seem to be reasonable to capture the aforementioned irregularities. For more elaborated explanation of reference measures one can refer to [35]. In the cases studied so far, these measures are taken to be either uniform or Gaussian.

Define $\overleftarrow{p_X}$ and $\overrightarrow{p_Y}$ as the images of $u_X$ and $u_Y$ under $f^{-1}$ and $f$ respectively. Then [20, 35] propose the following postulate for inferring the causal direction based on ICM:

**Postulate 1.** *If* $X \to Y$, *then*

$$D_{KL}(p_Y \| u_Y) = D_{KL}(p_X \| u_X) + D_{KL}(\overrightarrow{p_Y} \| u_Y) \tag{8}$$

Since this postulate is defined based on information geometric quantities, it is known as *Information Geometric Causal Inference (IGCI)*. It can be shown that whenever this postulate holds in one direction it cannot hold in the other direction. More precisely one has:

**Theorem 5.** *[20, Thm. 3] Let $f$ be non-trivial in the sense that the image of $u_X$ under $f$ does not coincide with $u_Y$. If eq. (8) holds, then one gets*

$$D_{KL}(p_X \| u_X) < D_{KL}(p_Y \| u_Y) + D_{KL}(\overleftarrow{p_X} \| u_X)$$

### 3.2.2 *ICM for Deterministic Linear High-dimensional Relations (Trace Condition)*

A causal inference framework for the study of multidimensional deterministic linear functions has been developed in [34, 77]. We briefly explain this framework here. Suppose that the data generating mechanism from a given input vector $\mathbf{X} \in \mathbb{R}^m$ is a matrix $A \in M_{m\times n}(\mathbb{R})$ and:

$$\mathbf{Y} = A\mathbf{X}. \tag{9}$$

Then the independence of cause and mechanism in such a relation has been interpreted as follows:

**Postulate 2. (Trace condition)** *In a linear functional relationship such as eq. (9) where $\mathbf{X}$ is the cause, the following equation holds approximately*

$$\tau(A\Sigma_X A^\top) = \tau(\Sigma_X)\tau(AA^\top),$$

*where $\tau(L)$ is the normalized trace for a given matrix $L \in M_{n\times n}(\mathbb{R})$, i.e.*

$$\tau(L) := \frac{\mathrm{tr}(L)}{n}$$

To investigate to what extent such an assumption holds [77] introduces $\Delta_{X\to Y}$ as follows

$$\Delta_{X\to Y} := \log \tau(A\Sigma_X A^\top) - \log \tau(\Sigma_X) - \log \tau(AA^\top) \tag{10}$$

$$\Delta_{Y\to X} := \log \tau(\Sigma_X) - \log \tau(\Sigma_Y) - \log \tau(A^{-1}A^{-\top}). \tag{11}$$

For a reason that will be clarified later we will make use of this expression for the case where $A \in M_{n\times n}(\mathbb{R})$ and as a result for such a $n$ we represent our expression with $\Delta^n_{X\to Y}$. In the next theorem which is one of the identifiability results derived for this ICM-based framework, $O(N)$ is the group of orthogonal matrices of order $N$.

**Theorem 6.** *[77]* **(CoM for finite dimensional linear relationships)** *Suppose $\Sigma$ is a given covariance matrix and suppose $A \in M_{n\times m}(\mathbb{R})$ is also a given matrix. Then if one generates $\Sigma_X = U\Sigma U^\top$ by uniformly choosing an orthogonal matrix $U$ from $O(n)$ then $\Sigma_X$ together with $A$, satisfies trace con-*

*dition in probability when* $n$ *tends to infinity. More precisely for a given* $\varepsilon$ *there exist* $\delta := 1 - \exp(\kappa(n-1)\varepsilon^2)$, $\kappa$ *being a constant where*

$$|\tau_m(A\Sigma_X A^\top) - \tau_n(\Sigma_X)\tau_m(AA^\top)| =$$
$$|\tau_m(AU\Sigma U^\top A^\top) - \tau_n(\Sigma)\tau_m(AA^\top)| \leqslant 2\varepsilon\|\Sigma\|\|AA^\top\|$$

*holds with probability* $\delta$.

In the above theorem (and the rest of the document) $\|.\|$ is the operator norm. The following lemma is a consequence of the previous theorem:

**Corollary 3.** *Suppose* $\Sigma$ *is a given covariance matrix and suppose* $A \in M_{n\times m}(\mathbb{R})$ *is also a given matrix. Then if one generates* $A_U = AU$ *by uniformly choosing an orthogonal matrix* $U$ *from* $O(n)$ *then* $A_U$ *together with* $\Sigma$*, satisfies trace condition in probability when* $n$ *tends to infinity More precisely for a given* $\varepsilon$ *there exist* $\delta := 1 - \exp(\kappa(n-1)\varepsilon^2)$, $\kappa$ *being a constant where*

$$|\tau_m(A_U\Sigma A_U^\top) - \tau_n(\Sigma_X)\tau_m(AA^\top)| =$$
$$|\tau_m(AU\Sigma U^\top A^\top) - \tau_n(\Sigma)\tau_m(AA^\top)| \leqslant 2\varepsilon\|\Sigma\|\|AA^\top\|$$

*holds with probability* $\delta$.

There is an important observation to make at this point; The fact that such a corollary holds was expected since independence is a mutual relationship; Selecting cause independently from mechanism would yield the same result as if one selects mechanism independently from the cause.

### 3.2.3 *IGCI and Trace Condition*

[35] proposes an argument that connects the trace condition to IGCI postulate when the underlying random variables have Gaussian distributions:

**Theorem 7.** *[35] Suppose* $X \sim \mathcal{N}(0, \Sigma_X)$ *is an* $N$ *dimensional random variable and* $Y = AX$ *where* $A \in M_{N\times N}(\mathbb{R})$*. Moreover suppose* $\mathcal{E}_X$ *and* $\mathcal{E}_Y$ *are manifold of isotropic Gaussian densities. Then the following relationship holds:*

$$D(P_Y\|\mathcal{E}_Y) = D(P_X\|\mathcal{E}_X) + D(\overrightarrow{P_Y}\|\mathcal{E}_Y) + \frac{N}{2}\Big[1 - \frac{\tau(\Sigma_X)\tau(AA^\top)}{\tau(\Sigma_{y_N})}\Big].$$

In other words this theorem states that orthogonality assumption is fullfiled for $X \to Y$ when the functional relationship between variables is linear, if and only if the trace condition is satisfied.

## 3.3 CAUSAL INFERENCE FOR TIME SERIES

The problem of causal inference in time series has been first formally studied by N. Wiener [72] and adapted in a practical form by C. Granger [27]. The causal inference postulate of Wiener-Granger causality states that a time series $X(t)$ causes a time series $Y(t)$ if the past of $X_t$ and $Y_t$ results in a better

prediction of the future of $Y_t$ than a prediction based on the past of $Y_t$ only. The mathematical formulation proposed by Granger was based on a linear model of vector autoregression [27]. Later generalization to other models have been proposed, including a non-parametric technique based on information theory, Transfer Entropy, introduced by T. Schreiber [60].

Although Wiener-Granger causality is by far the most popular approach for time series, other methods for inferring causal relations have been suggested by physicists to infer driver-response relationships in coupled dynamical systems. These techniques rely on the notion of dynamical interdependence [6, 22, 41, 58]. Measures of interdependence have been criticized for being affected by the difference in complexity of the trajectories of the underlying systems, possibly generating erroneous causal inferences [6, 50]. On the other hand neither methods based on Wiener-Granger causality nor dynamical interdependences can infer causal relations in complete absence of noise or chaotic behaviour.

Finally, in machine learning community, there has been other attempts to establish frameworks of causal inference for time series based on the generalization of available causal inference methods for static random variables in which time can be neglected. Two of these works are based on extensions of additive noise models for causal inference [19, 31]. Granger causality is unable to address nonlinear instantaneous causal relations and both of the aforementioned methods are unable to account for hidden variables. Motivated to address these shortcomings [48] proposes another causal inference method for time series which is based on restricted functional causal models called TiMiNo.

## 3.4 CAUSAL INFERENCE APPLIED TO NEURAL DATA

Unlike computers that rely highly on sequential processing, brain information processing is highly distributed among a large number of modules [44]. Since the anatomical connectivity between cortical modules is widespread and largely bidirectional, the detailed organization of information processing needs to be inferred from brain activity itself. In particular, finding measures to quantify the directed influence between several brain regions from signals recorded in each of them, called effective connectivity measures, has raised considerable interest in the last decade [13]. Effective connectivity measures essentially aim at inferring the directionality and strength of the interactions between different brain regions during a particular task or brain state. The purpose of these approaches is to help understand the underlying dynamical mechanisms of information processing [21] as well as improving the diagnosis and treatment of brain disease such as epilepsy [4, 18, 38, 55].

Effective connectivity studies based on functional Magnetic Resonance Imaging data can mainly be divided into two categories [68]; the ones re-

lying on Dynamic Causal Modelling (DCM) [23] and the others relying on Granger causality [69]. However, recent works also started to consider methods based on the causal Markov condition [52]. Effective connectivity studies based on electrical activity measurements, like Electroencephalography and Magnetoencephalography or local field potentials (LFPs), are either based on Granger causality [21] (and its generalizations such as transfer entropy [70], partial directed coherence [57, 59]) or rely on dynamical interdependence measures [41, 51, 58]. Although DCM has been used extensively in neuroimaging research, recently its effectiveness has been challenged [43]. Granger causality based methods, as previously mentioned, are unable to find the causal direction in absence of noise or chaotic behaviour. Moreover, Granger causality relies on a postulate involving predictability of the time series as stated by Wiener [72], while other postulates might be exploited advantageously.

# 4

# ICM FOR LINEAR SYSTEMS

In this chapter we introduce a causal inference framework for the study of causal inference for time series that are inputs and outputs of linear time invariant systems. We show that this framework is an extension of already established trace method in the limit. Then in the following chapter we derive some identifiability results special for this framework and finally we test our framework on synthesized and real world data. We start by explaining the idea through two examples. We then derive our formal expressions through these example and then compare them to the asymptotic behaviour of trace expressions introduced in [34].

## 4.1 DEMONSTRATIVE EXAMPLES

We first illustrate an intuitive justification of the framework through two simple examples. The first example is the case where the input of LTI is a circular process. The second example is the case where the input signal to the system is a white noise.

### 4.1.1 *Example: LTI with White Noise as Input*

For a given LTI with impulse response function $h_t$ suppose $\{X_t\}$ the input of the system is a white noise with $\sigma^2$ as its constant power spectrum. Based on proposition 2 we get

$$\forall \nu \quad -\frac{1}{2} \leqslant \nu \leqslant \frac{1}{2}, \quad S_{yy}(\nu) = S_{xx}(\nu)|\hat{h}(\nu)|^2 = \sigma^2|\hat{h}(\nu)|^2. \tag{12}$$

As one can see, transfer function modulates the power spectral density of the stochastic process by changing the contribution of each frequency to the total variance of input and by doing so generates the SDF of output signal or $S_{yy}$. Now suppose $\hat{h}_\nu$ is non-vanishing and therefore $h^{(-1)}$ exist. Since the output of this filter has a constant power in any frequency, $h^{(-1)}$ should have been "designed" in a way that modulates the power values larger than $\sigma^2$ and attenuates them to become $\sigma^2$ and it increases the power values with frequencies less than $\sigma^2$ to become $\sigma^2$. Therefore the backward filter is highly informative about the input signal $\{Y_t\}$ and this is suspicious relationship if we are ought to assume that the postulate of independence between cause and mechanism holds. An illustration of this situation has been depicted in fig. 1. This motivates us to define a measure of independence between power spectra of the input and output on one hand and the energy spectra of impulse response function on the other hand. We represent dependence of input and mechanism in terms of covariance between transfer

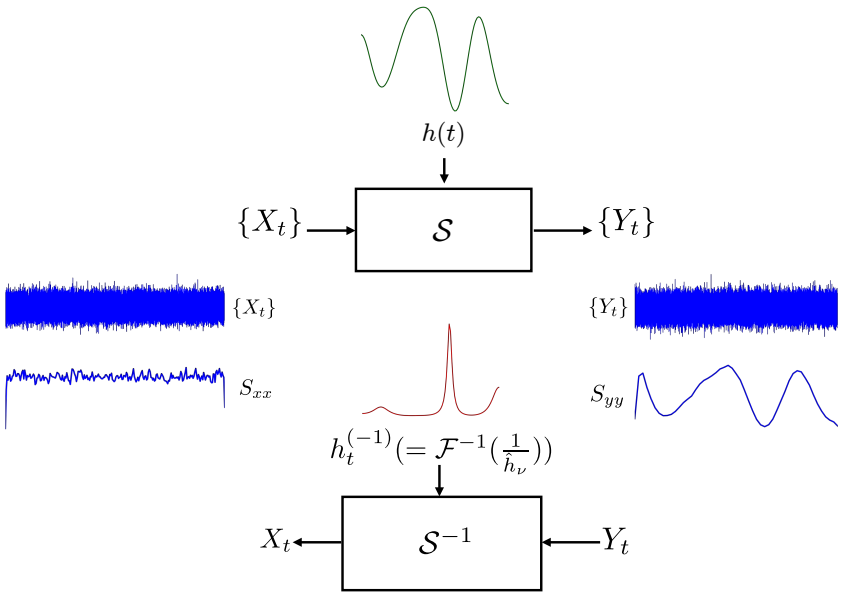



Figure 1: A schematic of a filter that takes as input a time series close to white noise (blue signal on the left). Therefore the spectral density of input $\{X_t\}$ is highly uncorrelated to $\hat{h}_\nu$. The output $\{Y_t\}$ as can be seen (on the right) has a spectral density very similar to the transfer function of $\{h_t\}$ (depicted in green). On the hand in the backward direction the transfer function have peaks at frequencies that the power spectrum of input ($\{Y_t\}$) has valleys. This makes the this transfer function and the spectral density $S_{yy}$ to have a highly negative correlation.

function and input power spectrum. More precisely, for $S_{xx}$ and $\hat{h}$ defined on $[-\frac{1}{2}, \frac{1}{2})$ and the uniform measure over $[-\frac{1}{2}, \frac{1}{2})$ as reference measure we get:

$$\mathbb{C}ov(S_{xx}, |\hat{h}|^2) = \mathbb{C}ov(S_{xx}, \frac{S_{yy}}{S_{xx}}) = \tag{13}$$
$$\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu - \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu)|^2 d\nu$$

Now based on eq. (12) we get

$$\mathbb{C}ov(S_{xx}, |\hat{h}|^2) = (\sigma^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu)|^2 d\nu - \sigma^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu')|^2 d\nu') = 0$$

### 4.1.2 *Example: Time Series With Finite Length*

Consider a linear time invariant (LTI) system with impulse response function $h_t$ with input and output $\{X_t\}$ and $\{Y_t\}$. More precisely for a given discrete zero mean wide sense stationary circular process the output of this filter is calculated by

$$\{X_t\}_{t=0}^{N-1} *_c \{h_t\}_{t=0}^{N-1} = \{Y_t\}_{t=0}^{N-1}.$$

In such a case applying the unitary DFT introduced will give us

$$\hat{Y}_\nu = \sqrt{N}\hat{X}_\nu \cdot \hat{h}_\nu$$

where $\{\hat{X}_\nu\}$ and $\{\hat{Y}_\nu\}$ are random processes in frequency domain. To derive a corollary based on [**?** ] we separate the real and imaginary part of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$. Then we get the following relation in real domain:

$$\hat{\mathbf{Y}}_R{:}\hat{\mathbf{Y}}_I = \sqrt{N}H \cdot (\hat{\mathbf{X}}_R{:}\hat{\mathbf{X}}_I) \tag{14}$$

Where $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_I$ are real and imaginary parts of $\mathbf{X}$ and $\hat{\mathbf{Y}}_R$ and $\hat{\mathbf{Y}}_I$ are real and imaginary parts of $\mathbf{Y}$ and

$$H = \left[\begin{array}{ccc|ccc} Re(\hat{h}_0) & & 0 & -Im(\hat{h}_0) & & 0 \\ & \ddots & & & \ddots & \\ 0 & & Re(\hat{h}_{N-1}) & 0 & & -Im(\hat{h}_{N-1}) \\ \hline Im(\hat{h}_0) & & 0 & Re(\hat{h}_0) & & 0 \\ & \ddots & & & \ddots & \\ 0 & & Im(\hat{h}_{N-1}) & 0 & & Re(\hat{h}_{N-1}) \end{array}\right]$$

and therefore a special case of ICM for linear deterministic highdimensional data introduced in section 3.2.2 will be retrieved. Matrix $H$ is fairly simple and an straightforward calculation shows:

$$\tau(HH^\top) = \Psi(\{\hat{\mathbf{h}}_\nu\}) = \Psi(\{h_t\}), \tag{15}$$

based on Plancherel theorem and definition of energy. Since DFT is a unitary transformation we have $\mathrm{tr}(\Sigma_X) = \mathrm{tr}(\Sigma_{\hat{X}})$ and $\mathrm{tr}(\Sigma_Y) = \mathrm{tr}(\Sigma_{\hat{Y}})$. Also from linearity of this transformation it follows that $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are zero mean processes. Moreover the following relationship holds for $\hat{\mathbf{X}}$ as a process:

$$\tau(\Sigma_{\hat{X}}) = \frac{1}{N}\sum_{f=0}^{N-1} \mathbb{C}ov(\hat{X}_f, \hat{X}_f^*) = \frac{1}{N}\sum_{f=0}^{N-1} \mathbb{E}(\hat{X}_f\hat{X}_f^*). \tag{16}$$

On the other hand, as explained above,

$$\mathrm{tr}(\Sigma_X) = \mathrm{tr}(\Sigma_{\hat{X}}) \Rightarrow \frac{1}{N}\sum_{f=0}^{N-1} \mathbb{E}_P(\hat{X}_f\hat{X}_f^*) = \frac{1}{N}\sum_{t=0}^{N-1} \mathbb{E}_P(X_tX_t^*) = \tag{17}$$

$$C_X(0) = \frac{1}{N}\sum_{f=0}^{N-1} S_{xx}[f] = P(\{X_t\}) \tag{18}$$

Same calculations can be carried out for $\Sigma_{\hat{Y}}$. This yields:

$$\Delta_{X\to Y} = \Delta_{\hat{X}\to\hat{Y}} = \log\tau(H\Sigma_{\hat{X}}H^\top) - \log\tau(\Sigma_{\hat{X}}) - \log\tau(HH^\top) = \log(P(\{Y_t\})) - \log(P(\{X_t\})) - \log(\Psi(\{\hat{h}_\nu\}))$$

The first equation in expression above follows again from the fact that DFT is a unitary transformation. And the last equation is based on eqs. (15) and (17). One needs to notice that this expression is nothing but the covariance expression in eq. (13) for discrete case.

As been discussed before there is a connection between two different methods of ICM and their relation has been emphasized through orthogonality of relative entropies [36]. For the sake of demonstration we also sketch the relation in this case to the information geometric representation of orthogonality, since later we will extend this result to the case of infinite time series and spectral densities. So far, for information geometric approach define 2N-dimensional isotropic Gaussian densities as reference measures for $X := \hat{\mathbf{X}}_R{:}\hat{\mathbf{X}}_I$ and $Y := \hat{\mathbf{Y}}_R{:}\hat{\mathbf{Y}}_I$ with equal means accordingly. By theorem 7 we get

$$D(P_Y\|\mathcal{E}_Y) = D(P_X\|\mathcal{E}_X) + D(\vec{P}_Y\|U_Y) + N\Big[1 - \frac{\tau(\Sigma_X)\tau(HH^\top)}{\tau(\Sigma_Y)}\Big] \tag{19}$$

Again based on eqs. (15) and (17) we get:

$$D(P_Y\|\mathcal{E}_Y) = D(P_X\|\mathcal{E}_X) + D(\vec{P}_Y\|U_Y) + N\Big[1 - \frac{P(X_t)\Psi(\{h_t\})}{P(Y_t)}\Big]$$

$$D(P_Y\|\mathcal{E}_Y) = D(P_X\|\mathcal{E}_X) + D(\vec{P}_Y\|U_Y) + N\Big[1 - \frac{\sum_{\nu=0}^{N-1} S_{xx}(\nu) \sum_{\nu=0}^{N-1} \frac{S_{yy}(\nu)}{S_{xx}(\nu)}}{\sum_{\nu=0}^{N-1} S_{yy}(\nu)}\Big]$$

Based on these observations we define our framework of causal inference for discrete LTI systems, i.e. LTI systems with discrete input and output.

## 4.2 SPECTRAL INDEPENDENCE CRITERIA (SIC)

The two examples in the previous section motivates the definition of a causal inference criteria based on spectral covariances. They also hint to a possible connection between the trace condition of [77] and the spectral relation since white noise example explained here is reminiscent of the example of linear relationship with an isotropic Gaussian distribution with identity covariance matrix as input presented as an example in [33]. We define our causal inference postulate as follows:

**Postulate 3. ( Spectral Independence Criteria)** *In a linear time invariant system with impulse response function* $h_t$ *and weakly stationary input and out-*

*put* $\{X_t\}$ *and* $\{Y_t\}$ *we say* $\{X_t\}$ *is the cause and* $\{Y_t\}$ *is the effect if the following equality approximately holds:*

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{S_{yy}(\nu)}{S_{xx}(\nu)}d\nu$$

To asses to what degree such a relation holds we introduce scale invariant expression $\Delta^{\infty}_{X_t \to Y_t}$ and for the backward direction $\Delta^{\infty}_{Y_t \to X_t}$ as follows:

$$\Delta^{\infty}_{X_t \to Y_t} := \log \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu - \log \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{S_{yy}(\nu)}{S_{xx}(\nu)}d\nu \quad (20)$$

$$\Delta^{\infty}_{Y_t \to X_t} := \log \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu - \log \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{S_{xx}(\nu)}{S_{yy}(\nu)}d\nu \quad (21)$$

Or written in terms of total power and energy:

$$\Delta^{\infty}_{X_t \to Y_t} = \log(P(\{Y_t\})) - \log(P(\{X_t\}) - \log(\Psi(\{h_t\}))$$
$$\Delta^{\infty}_{Y_t \to X_t} = \log(P(\{X_t\})) - \log(P(\{Y_t\}) - \log(\Psi(\{h_t^{(-1)}\}))$$

By scale invariance in the above expression what we mean is that if the input signal is multiplied by a nonzero value (whether in time or frequency domain) the value of the expression will not change.

**Remark 7.** *For the case of circular processes, every summand in eqs.* (20) *and* (21) *should be stated in terms of discrete Fourier transform; this means that integrals should be replaced by sums over frequency domain.*

### 4.2.1 *Relation between SIC and Trace Condition*

In this section we consider the asymptotic behaviour of trace condition [34] used to analyse the linear relationship arose between a truncated version of the input and output time series of an LTI and show that in the limit of sample size the trace expressions eqs. (10) and (11) approach to spectral expressions in eqs. (20) and (21).

As before suppose $\{X_t\}$ and a linear filter $\mathcal{S}$ are given. Then by proposition 1 there exists a impulse response function $h_t$ where

$$\forall \tau \in \mathbb{Z}, \quad Y_\tau = (\{X_t\} * \{h_t\})(\tau). \quad (22)$$

Lets take only a finite length of this convolution into account and write the following equation as an approximation for eq. (22):

$$Y'_t = \sum_{k=0}^{2N-1} X_{t-k} h_k \quad (23)$$

But this on the other hand, when written for the $2N$ elements of $\{Y'_t\}$ around index $0$ gives the following matrix relationship:

$$\begin{bmatrix} Y'_{-N} \\ Y'_{-N+1} \\ \vdots \\ Y'_{N-2} \\ Y'_{N-1} \end{bmatrix} = \begin{bmatrix} h_0 & h_{-1} & \cdots & h_{-2N+1} \\ h_1 & h_0 & \cdots & h_{-2N+2} \\ & \vdots & & \\ h_{2N-2} & h_{2N-3} & \cdots & h_{-1} \\ h_{2N-1} & h_{2N-2} & \cdots & h_0 \end{bmatrix} \begin{bmatrix} X_{-N} \\ X_{-N+1} \\ \vdots \\ X_{N-2} \\ X_{N-1} \end{bmatrix}. \tag{24}$$

If we name the vector on the left as $\mathbf{y}_N$, the matrix as $H^N$ and the right vector as $\mathbf{x}_N$ then the trace expression yields:

$$\Delta^{2N}_{X\to Y} = \log(\Sigma_{\mathbf{y}_N}) - \log(\Sigma_{\mathbf{x}_N}) - \log(H^N H^{N\top}) \tag{25}$$

Define $T_N := \tau(H^N H^{N\top})$. Now we show that $T_N$ converges to $\Psi(\{h_t\}) = \sum_{k=-\infty}^{\infty} |h_k|^2$, when the latter exists.

**Lemma 4.** *Suppose $h_t$ is absolutely convergent in the square norm. Then $T_N$ converges to $\Psi(\{h_t\})$.*

*Proof.* First lets simplify the expression for $T_N$:

$$\begin{aligned} T_N := \tau(H^N H^{N\top}) = \frac{1}{2N}\sum_{i,j}[H^N]^2_{ij} &= \sum_{k=-2N+1}^{2N-1} |h_k|^2 \frac{2N-|k|}{2N} \\ &= \sum_{k=-2N+1}^{-1} |h_k|^2 \frac{2N-|k|}{2N} + \sum_{k=0}^{2N-1} |h_k|^2 \frac{2N-|k|}{2N}. \end{aligned} \tag{26}$$

Its easy to see that $T_N$ is an increasing sequence of $N$. Moreover it is bounded by $\sum_{-\infty}^{\infty} |h_k|^2 < \infty$. Therefore this series converges. In order to show that it converges to $\Psi(\{h_t\})$, we first notice that for a given $\varepsilon$, there exist $m_0 \in \mathbb{N}$ such that

$$\forall m > m_0 \quad \Big| \sum_{k=-m}^{m} |h_k|^2 - \Psi(\{h_t\})\Big| < \varepsilon. \tag{27}$$

Now take $N_{m_0} > \frac{m_0 2^{m_0+1}|h_{m_0}|^2}{\varepsilon}$. We have

$$N_{m_0} > \frac{m_0 2^{m_0+1}|h_{m_0}|^2}{\varepsilon} \Rightarrow \frac{|h_{m_0}|^2 m_0}{2N_{m_0}} < \frac{\varepsilon}{2^{m_0+2}}.$$

Same can be done for any $0 \leqslant k \leqslant m_0$, i.e. there exist $N_k$ such that:

$$\frac{|h_k|^2 k}{2N_k} < \frac{\varepsilon}{2^{k+2}}$$

Now take $N_{max} = \max\{N_0, N_1, ..., N_{m_0}\} + 1$. Then obviously we get:

$$||h_k|^2 - \frac{|h_k|^2(2N_{max} - k)}{2N_{max}}| < \frac{\varepsilon}{2^{k+2}}$$

And therefore:

$$\sum_{k=0}^{m_0} ||h_k|^2 - \frac{|h_k|^2(2N_{max} - k)}{2N_{max}}| < \sum_{k=0}^{m_0} \frac{\varepsilon}{2^{k+2}} < \frac{\varepsilon}{2} \tag{28}$$

Similar results hold for the first sum term in eq. (26) and by taking the maximum of two $N_{max}$'s (say $N'_{max}$) and considering the fact that $T_N$ is increasing and by the application of triangular inequality for eq. (27), we can easily infer that

$$\forall N > N'_{max} \quad |T_N - \Psi(\{h_t\})| < \varepsilon.$$

□

**Remark 8.** *[25, pp. 378] When* $\sum_{k=-\infty}^{\infty} |h_k|^2 < \infty$*, the LTI* $\mathcal{L}$ *is stable in the sense that there exist* $A > 0$ *such that* $\|\mathcal{L}(\{x_t\})\|_\infty \leqslant A\|x\|_\infty$ *for any* $\{x_t\} \in l^1(\mathbb{R})$.

We also need to prove that $Y'_k$'s in eq. (24) are asymptotically converging to $Y_k$'s in the following sense:

**Lemma 5.** *Suppose an LTI filter* $\mathcal{S}$ *with zero mean weakly stationary processes as input* ($\{X_t\}$) *and output* ($\{Y_t\}$) *has been given. Then eq.* (22) *holds and therefore we can get a truncated linear relationship as eq.* (24)*. If* $\{h_t\}$ *would be absolutely convergent then we get the following relationship between covariance matrices of* $Y_{-N:N-1}$ *and* $Y'_{-N:N-1}$*:*

$$\lim_{N\to\infty} |\tau(\Sigma_{Y_{-N:N-1}}) - \tau(\Sigma_{Y'_{-N:N-1}})| = 0,$$

*Proof.* For simplicity of calculations we name $2N$ dimensional random vectors $Y'_{-N:N-1}$ and $Y_{-N:N-1}$ as $\mathbf{Y}'$ and $\mathbf{Y}$ and their covariance matrices with $\Sigma_{Y'}$ and $\Sigma_Y$ respectively. Then we have:

$$|\tau(\Sigma_{Y_{-N:N-1}}) - \tau(\Sigma_{Y'_{-N:N-1}})| = |\tau(\mathbb{E}_P(YY^\top)) - \tau(\mathbb{E}_P(Y'Y'^\top))| \overset{*}{=}$$
$$\frac{1}{2N}|\mathbb{E}_P(Y^\top Y) - \mathbb{E}_P(Y'^\top Y')| = \frac{1}{2N}|\mathbb{E}_P\big((Y-Y')^\top(Y+Y')\big)| \leqslant$$
$$\frac{1}{2N}\mathbb{E}_P|(Y-Y')^\top(Y+Y')| \leqslant$$
$$\frac{1}{2N}\mathbb{E}_P\Big(\sqrt{(Y-Y')^\top(Y-Y')}\sqrt{(Y+Y')^\top(Y+Y')}\Big) \leqslant$$
$$\frac{1}{2N}\sqrt{\mathbb{E}_P((Y-Y')^\top(Y-Y'))}\sqrt{\mathbb{E}_P((Y+Y')^\top(Y+Y'))} =$$
$$\sqrt{\frac{1}{2N}\mathbb{E}_P((Y-Y')^\top(Y-Y'))}\sqrt{\frac{1}{2N}\mathbb{E}_P((Y+Y')^\top(Y+Y'))} \overset{**}{=}$$
$$\sqrt{\tau(\Sigma_{Y-Y'})}\sqrt{\tau(\Sigma_{Y+Y'})}$$

where (*) and (**) follows from the fact that one can take trace (or normalized trace) into expectation and vice versa, and moreover from the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any two matrices that their multiplication is well defined. The inequalities are the result of the application of Cauchy-Schwartz inequality for covariances of random variables. First we show that $\sqrt{\tau(\Sigma_{Y+Y'})}$ is bounded as a function of $N$. Define $\{h_t^{(j)}\}$ as follows

$$h_t^{(j)} = \begin{cases} 2h_t & \text{if } -N \leqslant t+j \leqslant N-1 \\ h_t & \text{otherwise} \end{cases}.$$

We can bound each element of diagonal of $\Sigma_{Y+Y'}$ as follows

$$[\Sigma_{Y+Y'}]_{jj} = \mathbb{E}_P\Big[(Y_j + Y'_j)^2\Big] = \mathbb{E}_P\Big[(\sum_{l=-\infty}^{\infty} X_{j-l}h_l^{(j)})^2\Big] \leqslant$$
$$\mathbb{E}_P\Big[(\sum_{l=-\infty}^{\infty} |X_{j-l}||h_l^{(j)}|)^2\Big] \leqslant 4\mathbb{E}_P\Big[(\sum_{l=-\infty}^{\infty} |X_{j-l}||h_l|)^2\Big] = 4C_Y(0),$$

and therefore $\tau(\Sigma_{Y+Y'})$ is bounded.

Now we show that each element of diagonal of $\Sigma_{Y-Y'}$ tends to zero when $N$ tends to infinity which will complete the proof. With overload of notation, in this case define $\{h_t^{(j)}\}$ as follows

$$h_t^{(j)} = \begin{cases} 0 & \text{if } -N \leqslant t+j \leqslant N-1 \\ h_t & \text{otherwise.} \end{cases}$$

Then for the $j$-th element of $\Sigma_{Y-Y'}$ we have

$$[\Sigma_{Y-Y'}]_{jj} = \mathbb{E}_P\Big[(Y_j - Y'_j)^2\Big] = \mathbb{E}_P\Big[(\sum_{l=-\infty}^{\infty} X_{j-l}h_l^{(j)})^2\Big] = \mathbb{E}_P\Big[(\sum_{\substack{l\geqslant N-j\\ l<-N-j}} X_{j-l}h_l)^2\Big]$$

Since autocorrelation function attains its maximum at $t = 0$ and

$$\forall i, j \in \mathbb{Z}, \quad \mathbb{E}_P(X_iX_j) \leqslant \sqrt{\mathbb{E}_P(X_i^2)\mathbb{E}_P(X_j^2)}$$

we get:

$$\forall i, j \in \mathbb{Z}, \quad \mathbb{E}_P(X_iX_j) \leqslant \mathbb{E}_P(X_0^2)$$

$$[\Sigma_{Y-Y'}]_{jj} = \mathbb{E}_P\Big[(\sum_{\substack{l\geqslant N-j\\ l<-N-j}} X_{j-l}h_l)^2\Big] \leqslant \sum_{\substack{l,l'\geqslant N-j\\ l,l'<-N-j}} \mathbb{E}_P(X_0^2)h_lh_{l'} =$$

$$\mathbb{E}_P(X_0^2)\sum_{\substack{l,l'\geqslant N-j\\ l,l'<-N-j}} h_lh_{l'} \leqslant \mathbb{E}_P(X_0^2)(\sum_{\substack{l\geqslant N-j\\ l<-N-j}} h_l)^2 \leqslant \mathbb{E}_P(X_0^2)(\sum_{\substack{l\geqslant N-j\\ l<-N-j}} |h_l|)^2$$

Now since $\{h_t\}$ is absolutely convergent, it follows that $[\Sigma_{Y-Y'}]_{jj}$ can be arbitrarily reduced by increasing $N$. Then it follows that $\tau(\Sigma_{Y-Y'})$ approaches to zero when $N$ tends to infinity. □

Finally to prove the theorem regarding the asymptotic behaviour of of trace method and its equivalence to SIC, we need one of the convergence theorems due to Szegö:

**Theorem 8. (Szegö's convergence theorem)***[28] Let $f : [-\frac{1}{2}, \frac{1}{2}) \to \mathbb{R}$ $f \in L^1$ be a bounded function and suppose $t_k$ is its Fourier series coefficients, i.e.*

$$t_k = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\nu)e^{ik\nu}d\nu, \quad t \in \mathbb{Z}.$$

*Consider Toeplitz matrices $T_n$ defined as*

$$[T_n]_{ij} = t_{i-j} \quad i, j \in \{0, ..., n-1\}$$

*with eigenvalues $\tau_{n,k}(0 \leqslant k \leqslant n-1)$. Then if $T_n$'s are Hermitian, i.e. $t_i = \bar{t_i}$ for any $i$, then for any continuous function $F$ we have:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} F(\tau_{n,k}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} F(f(\nu))d\nu$$

We are ready to state our convergence theorem:

**Theorem 9.** *For a given windowed discrete time series, elements of trace condition, i.e. $\Delta^n_{X\to Y}$ and $\Delta^n_{Y\to X}$, asymptotically (increasing the size of window as*

*defined in eq. (24)) approach to the spectral values of time series on infinite domain. As a result the spectral density based estimator coincides with the trace based estimator in the limit, and more precisely*

$$\lim_{N\to\infty} \tau(\Sigma_{x_N}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu, \quad \lim_{N\to\infty} \tau(\Sigma_{y_N}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu,$$

$$\text{and} \quad \lim_{N\to\infty} T_N = \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu)|^2 d\nu,$$

*where $T_N$ is defined as in eq. (26). And eventually:*

$$\lim_{n\to\infty} \Delta^n_{X\to Y} = \Delta^\infty_{X\to Y} \quad \lim_{n\to\infty} \Delta^n_{Y\to X} = \Delta^\infty_{Y\to X}$$

*Proof.* Both $\Sigma_{x_N}$ and $\Sigma_{y_N}$ are hermitian Toeplitz matrices and based on theorem 8 where F has been chosen as identity function and also applying lemma 5 we get:

$$\lim_{N\to\infty} \tau(\Sigma_{x_N}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(\nu)d\nu \tag{29}$$

$$\lim_{N\to\infty} \tau(\Sigma_{y_N}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(\nu)d\nu \tag{30}$$

Moreover by Plancherel's theorem and lemma 4 it follows that:

$$\lim_{N\to\infty} T_N = \Psi(\{h_t\}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu)|^2 d\nu \tag{31}$$

□

This theorem shows that trace expressions calculated for windowed version of time series are nothing but estimates of spectral expression and therefore justifies that these two different methods for causal inference are indeed consistent with each other. In the next chapter we establish a different connection (based on information theory) to the another ICM-based causal inference methods for nonlinear relationships [20] when the time series are discrete weakly stationary Gaussian processes.

# 5

# SIC AND IGCI FOR GAUSSIAN PROCESSES

In section 3.2.3 we have described that [35] established a connection between the causal inference method for linear high dimensional data and IGCI method for nonlinear functions. Later we applied this derivation on circular processes to illustrate an equivalent information geometric relation in therms of spectral densities for weakly stationary Gaussian circular processes. In this section we extend this result to the case of discrete weakly stationary and purely non-deterministic Gaussian processes. First we start by finding the isotropic Gaussian process (Gaussian white noise) which minimized the relative entropy rate of a given Gaussian process with respect to it. Name the set of all isotropic Gaussians defined on the same probability space of $\{X_t\}$ as $\mathcal{E}_{X_t}$ where we parametrise the elements of this family with $U_t^\sigma$ and $\sigma$ represents the constant power spectral density of the white noise. Then we have the following

**Lemma 6.** *Suppose $\{X_t\}$ is a zero mean purely nondeterministic weakly stationary Gaussian processes with SDF's $S_{xx}$ and take $\mathcal{E}_{X_t}$ to be the set of discrete isotropic weakly stationary Gaussian processes defined on the same probability space. Then:*

$$\bar{D}(P_{X_t}\|\mathcal{E}_{X_t}) = -\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left(\frac{S_{xx}(\nu)}{P(X_t)}\right) d\nu \tag{32}$$

*Proof.* In order to find the isotropic Gaussian process with the minimum distance, we take the derivative of $\bar{D}(P_{X_t}\|P_{U_t^\sigma})$ and look for its singular value. We do this calculation by means of lemma 1 (One needs to notice that we can use this lemma because the condition $(ii)$ is satisfied):

$$\frac{d\bar{D}(P_{X_t}\|P_{U_t^\sigma})}{d\sigma} = \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(-\frac{f(\nu)}{\sigma^2} + \frac{f(\nu)}{\sigma^2}\frac{\sigma}{f(\nu)}\right) d\nu = 0 \Rightarrow$$
$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{f(\nu)}{\sigma^2} = \frac{1}{\sigma} \Rightarrow \sigma = P(X_t)$$

And we get:

$$\bar{D}(P_{X_t}\|\mathcal{E}_{X_t}) = \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{f(\nu)}{P(X_t)} - 1 - \ln\left(\frac{f(\nu)}{P(X_t)}\right) d\nu.$$

Using the definition of P we get:

$$\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{f(\nu)}{P(X_t)} - 1 - \ln\left(\frac{f(\nu)}{P(X_t)}\right) d\nu =$$

$$\frac{1}{2}\frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} f(\nu)}{P(X_t)} - \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} 1 - \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left(\frac{f(\nu)}{P(X_t)}\right) d\nu = -\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left(\frac{f(\nu)}{P(X_t)}\right) d\nu$$

And this proves the argument. □

Now based on this observation we calculate all the terms that are calculated in section 4.5 of [35]. Here $\mathcal{E}_{X_t}$ and $\mathcal{E}_{Y_t}$ represent the set of Gaussian white noise processes defined on the same probability space as $\{X_t\}$ and $\{Y_t\}$ respectively.

**Theorem 10.** *Suppose weakly stationary Gaussian process $\{X_t\}$ and $\{Y_t\}$ are the input and output for an LTI system $\mathcal{S}$ respectively and $h_t$ is the impulse response function of $\mathcal{S}$. If $\hat{h}_\nu$ satisfies any of the conditions of lemma 1 when replaced with $S_{yy}$, then*

$$\bar{D}(P_{Y_t}||\mathcal{E}_{Y_t}) = \bar{D}(P_{X_t}||\mathcal{E}_{X_t}) + \bar{D}(\overrightarrow{P_{Y_t}}||U_{Y_t}) + \frac{1}{2}\left(1 - \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{S_{yy}}{S_{xx}}}{\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}}\right)$$

*Proof.* Using lemma 6 we have

$$\bar{D}(P_{X_t}||\mathcal{E}_{X_t}) = -\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left(\frac{S_{xx}(\nu)}{P(X_t)}\right) d\nu$$

$$\bar{D}(P_{Y_t}||\mathcal{E}_{Y_t}) = -\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left(\frac{S_{yy}(\nu)}{P(Y_t)}\right) d\nu$$

Transforming $\{U_{X_t}\}$ and $\{h_t\}$ to Fourier domain its easy to see that $\{\overrightarrow{Y_t}\}$ is a zero mean weakly stationary Gaussian process with SDF $P(X_t)|\hat{h}(\nu)|^2$ according to proposition 2. Therefore using lemma 1 we get

$$\bar{D}(\overrightarrow{P_{Y_t}}||U_{Y_t}) = \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P(X_t)|\hat{h}(\nu)|^2}{P(Y_t)} - 1 - \log\frac{P(X_t)|\hat{h}(\nu)|^2}{P(Y_t)}\right) d\nu$$

which completes the proof. □

This theorem now shows that the orthogonality in information space introduced as a criteria for causal inference in [20, 35] in the case of discrete weakly stationary zero mean Gaussian processes where the function is an LTI filter is equivalent to the independence condition introduced for spectral independence criterion.

**Remark 9.** *Similar to the information geometric approach in [20, 35], its possible to interpret these results in terms of information geometry if one takes*

*$\mathcal{E}_{X_t}$ and $\mathcal{E}_{Y_t}$ as manifold of isotropic Gaussian processes. One needs to know that this type of infinite dimensional exponential manifolds are well defined and they inherit the information geometric properties of the manifolds of multidimensional exponential distributions [49].*

Now that we showed the close relations between SIC and other methods based on ICM we derive some identifiability results for SIC method in the next chapter.

# 6

# IDENTIFIABILITY RESULTS

It has been argued in [77] that $\Delta^{n}_{X\to Y}$ is not necessarily positive or negative in correct causal direction. Rather the causal inference framework established for numerical tests is based on the fact that $\Delta^{n}_{X\to Y}$ expression has smaller absolute value in correct direction comparing to the backward direction. This argument is easy to illustrate for $\Delta^{\infty}_{X\to Y}$ as well; just consider LTI filter with white noise as input where in one case the transfer function is constant but greater than one and in the other case its constant and smaller than one. Our claim still is that the spectral expression $\Delta^{\infty}_{X\to Y}$ is closer to zero in absolute value when $X \to Y$. The following section is composed of concentration of measure arguments justifying this point.

## 6.1 CONCENTRATION OF MEASURE (COM)

Here we state and prove two different concentration of measure theorems to indicate that the delta expressions eqs. (20) and (21) yield smaller values (in absolute sense) in the correct causal direction when some conditions hold for the data generating model. We first assert a concentration of measure theorem for FIR filters. Before stating the theorem we state a consequence of a limit theorem due to Szegö which will be used in the proof of the theorem:

**Lemma 7.** *[28, corr. 4.2] As before (c.f. to theorem 8) suppose $T_n$ are sequences of Toeplitz matrices and with eigenvalues $\tau_{n,i}$ associated with SPD $g$ and Fourier coefficients of $g$, $t_i$ are absolutely summable. Then*

$$\lim_{n\to\infty} \max_{i} \tau_{n,i} = \max_{x\in[-\frac{1}{2},\frac{1}{2})} g(x)$$

The next theorem shows that for any given weakly stationary input when the forward coefficients of an FIR filter are chosen based on a rotation invariant prior the spectral independence criterion will be satisfied with high probability when the order of the filter increases.

**Theorem 11. (CoM for FIR filters)** *Suppose $b_i$'s ($0 \leqslant i \leqslant m-1$) are given real numbers. Define $\mathcal{S}$ to be the FIR filter with coefficients $b_i$ and impulse response function $h$. Now suppose $U \in O(m)$ has been chosen uniformly. Suppose $\mathcal{S}^{U}$ is another FIR with coefficients $U^{\top}\boldsymbol{b}$ and impulse response function $h^{U}$. Then for a weakly stationary input $\{X_t\}$ where $C_X(\tau)$ is absolutely summable and for a given $\epsilon$,*

$$\frac{|\int S^{U}_{yy}(\nu)d\nu - \Psi(\mathbf{b})P(X_t)|}{\Psi(\mathbf{b})\max_{\nu} S_{xx}(\nu)} = \frac{|\int S_{xx}(\nu)|\widehat{h^{U}}(\nu)|^2 d\nu - \Psi(\mathbf{b})P(X_t)|}{\Psi(\mathbf{b})\max_{\nu} S_{xx}(\nu)} \leqslant 2\epsilon$$

*with probability* $\delta := 1 - \exp(\kappa(m-1)\varepsilon^2)$ *where again* $\kappa$ *is a constant.*

*Proof.* Without loss of generality and for the sake of simplicity we only consider the positive indices of the time series and we take the filter to be causal; other cases can be treated in the similar way. Then the following relation holds between input and output of the filter:

$$\forall i, \quad 0 \leqslant i \leqslant N-1 \quad Y_i = \sum_{j=0}^{m-1} b_j X_{i-j}$$

Formulated in terms of matrices the above relation can be represented as

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{N-2} \\ Y_{N-1} \end{bmatrix} = \begin{bmatrix} b_{m-1} & b_{m-2} & \cdots & b_0 & 0 & \cdots & 0 & 0 \\ 0 & b_{m-1} & \cdots & b_1 & b_0 & \cdots & 0 & 0 \\ & & \ddots & & & & & \\ 0 & 0 & \cdots & b_{m-1} & \cdots & b_1 & b_0 & 0 \\ 0 & 0 & \cdots & 0 & b_{m-1} & \cdots & b_1 & b_0 \end{bmatrix} \begin{bmatrix} X_{-m+1} \\ X_{-m+2} \\ \vdots \\ X_{N-2} \\ X_{N-1} \end{bmatrix},$$

where we call the above $N \times (N+m-1)$ matrix as $B$. We define $\Sigma_X^i \in M_{m\times m}(\mathbb{R})$ to be the covariance matrices as follows:

$$\forall i \quad 0 \leqslant i \leqslant N-1 \quad 0 \leqslant j,k \leqslant m-1 \quad [\Sigma_X^i]_{jk} = \mathbb{C}\mathrm{ov}(X_{i+j}, X_{i+k})$$

If we take $\Sigma_{X_{0:N-1}}, \Sigma_{Y_{0:N-1}} \in M_{N\times N}(\mathbb{R})$ to be the covariance matrices for $X_{0:N-1}$ and $Y_{0:N-1}$ respectively, then we have

$$\Sigma_{Y_{0:N-1}} = B\Sigma_{X_{-m+1:N-1}} B^\top$$

Also define $\Sigma^U_{Y_{0:N-1}}$ to be the covariance matrix of the output for FIR $\mathcal{S}'$ with $\mathbf{b}' = U^\top \mathbf{b}$. also assume the spectrum of the output for this filter is $S^U_{yy}$. One can write diagonal elements of $\Sigma_{Y_{0:N-1}}$ and $\Sigma^U_{Y_{0:N-1}}$ based on the above equation as follows:

$$[\Sigma_{Y_{0:N-1}}]_{ii} = \mathbf{b}^\top \Sigma_X^i \mathbf{b}, \quad [\Sigma^U_{Y_{0:N-1}}]_{ii} = \mathbf{b}^\top U \Sigma_X^i U^\top \mathbf{b}$$

and therefore the normalized trace of $\Sigma_{Y_{0:N-1}}$ can be written as

$$\tau_N(\Sigma_{Y_{0:N-1}}) = \frac{1}{N}\mathbf{b}^\top \sum_{i=0}^{N-1} \Sigma_X^i \mathbf{b}, \quad \tau_N(\Sigma^U_{Y_{0:N-1}}) = \frac{1}{N}\mathbf{b}^\top U \sum_{i=-m+1}^{N-1} \Sigma_X^i U^\top \mathbf{b}$$

Define $\Sigma := \frac{1}{N}\sum_{i=-m+1}^{N-1} \Sigma_X^i$. Taking $A = \mathbf{b}^\top$ in corollary 3 in chapter 3 for a randomly selected $U$ we get:

$$|\frac{1}{N}\mathbf{b}^\top U \Sigma U^\top \mathbf{b} - \frac{1}{N}\tau_m(\Sigma)\langle \mathbf{b}, \mathbf{b}\rangle| \leqslant 2\varepsilon \|\Sigma\| \langle \mathbf{b}, \mathbf{b}\rangle$$

And therefore

$$|\tau_N(\Sigma^U_{Y_{0:N-1}}) - \frac{1}{N}\tau_m(\Sigma)\langle \mathbf{b}, \mathbf{b}\rangle| \leqslant 2\varepsilon\|\Sigma\|\langle \mathbf{b}, \mathbf{b}\rangle$$

On the other hand the elements of diagonals of $\Sigma^i_X$'s are $C_X(0)$. Therefore:

$$\frac{1}{N}\tau_m(\Sigma) = \frac{m(N)C_X(0)}{mN} = P(\{X_t\})$$

Since $\Sigma^i_X$'s are principal submatrices of $\Sigma_{X_{0:N-1}}$ therefore by corollary 2 in chapter 2 we get

$$\rho(\Sigma) = \|\Sigma\| = \|\frac{1}{N}\sum_{i=0}^{N-1}\Sigma^i_X\| \leqslant \frac{1}{N}\sum_{i=0}^{N-1}\|\Sigma^i_X\| \leqslant \rho(\Sigma_{X_{0:N-1}}).$$

Now since $C_X(\tau)$'s are absolutely summable we apply lemma 7 and we get

$$\rho(\Sigma_{X_{0:N-1}}) \leqslant \max_\nu S_{xx}(\nu)$$

And this together with theorem 8 completes the proof. □

**Remark 10.** *In the preceding lemma for weakly stationary input, $\Sigma^i_X$'s and $\Sigma^i_Y$'s will be independent of $i$ (they will be all equal). We believe the assumption of weakly stationarity of input is strong and one can derive a similar result for cyclostationary processes [24], but we leave such an extension to future work.*

The second theorem is only confined to the case where the time series is finite. For this theorem we introduce some new notations. In what follows we define an action of $O(N)$ over the set of orthogonal matrices in $\mathbb{R}$ that is not compatible to group. For $U \in O(N)$ and finite time series with power spectrum $S_{xx}$ define $U \bullet S_{xx} := \mathrm{diag}(UD(S_{xx})U^\top)$ where $D(S_{xx})$ is a diagonal matrix with diagonal entries being elements of $S_{xx}$ in order. In other words $[U \bullet S_{xx}]_{ii} = \sum_{j=1}^{N}|u_{ij}|^2[S_{xx}]_i$. We will also need the following lemma:

**Lemma 8.** *Suppose $\Sigma \in M_{N\times N}(\mathbb{R})$ and $A \in M_{N\times N}(\mathbb{R})$ are given diagonal matrices. If $U$ has been uniformly selected from $O(N)$ in $\mathbb{R}$,then for a given $\varepsilon$ there exist $\delta := 1 - \exp(\kappa(2N-1)\varepsilon^2)$, $\kappa$ being a constant where*

$$|\tau_N(A\ \mathrm{diag}(U\Sigma U^\top)A^\top) - \tau_N(\Sigma_X)\tau_N(AA^\top)| = \leqslant 2\varepsilon\|\Sigma\|\|AA^\top\|$$

*Proof.* Since we have

$$|\tau_{2N}(AU\Sigma U^\top A^\top) - \tau_{2N}(\Sigma)\tau_{2N}(AA^\top)| = |\tau_N(A\ \mathrm{diag}(U\Sigma U^\top)A^\top) - \tau_N(\Sigma_X)\tau_N(AA^\top)|,$$

applying theorem 6 proves the lemma. □

Based on this definition we can state and prove our concentration of measure theorem for finite time series with mixing spectrum (as to be explained below):

**Theorem 12. (CoM for finite time series)** *Suppose* $U$ *has been randomly selected from the Haar measure over orthogonal group of* $M_{N\times N}(\mathbb{R})$ *and* $\{X_t\}$ *is a circular process. Then the following holds*

$$\left|\sum |\hat{h}(\nu)|^2 (U \cdot S_{xx})(\nu) - \sum S_{xx}(\nu) \sum |\hat{h}(\nu)|^2\right| = \\ \leqslant 2\varepsilon \max_{\nu}(S_{xx}(\nu)) \max_{\nu}(|\hat{h}(\nu)|^2)$$

*with probability* $\delta := 1 - \exp(\kappa(2N-1)\varepsilon^2)$ *where again* $\kappa$ *is a constant.*

*Proof.* The proof follows from the definition of $(\bullet)$, the fact that operator norm for diagonal matrices is equal to the largest diagonal value and an application of lemma 8.

□

The reason that we call these time series as time series with mixing spectrum is that for any given $S_{xx}$, $U \cdot S_{xx}$ is another spectrum that each element of it is a weighed average of the spectrum of $S_{xx}$ for any possible weighting. Although there is no practical justification for existence of processes with such priors over their spectra we found it worthwhile to mention this established result.

Since in this section we showed that under some assumptions the spectral estimator can arbitrarily get close to zero in the right direction when the order of the filter increases we show in the next section that it cannot be the case that the spectral estimator to get arbitrarily close to zero in both directions; this will complete our causal inference framework in terms of its identifiability strength.

## 6.2 VIOLATION OF SIC

Finally, we are ready to present our violation of spectral formula condition; If SIC is satisfied in forward direction $(X \to Y)$, i.e. $\Delta^{\infty}_{X\to Y}$ is close to zero in absolute value then it is violated in the backward direction, i.e. $\Delta^{\infty}_{Y\to X}$ cannot be close to zero an absolute value either.

**Lemma 9. (Violation of SIC)** *For a given linear filter, under mild conditions, the following relationship holds*

$$\Delta^{\infty}_{X\to Y} + \Delta^{\infty}_{Y\to X} = -\log\left(1 - \mathbb{C}\mathrm{ov}(\hat{h}^2, \frac{1}{\hat{h}^2})\right)$$

*Proof.* Based on definitions eqs. (20) and (21) we have:

$$\Delta^{\infty}_{X\to Y} = \log(\int S_{yy}(\nu)d\nu) - \log(\int \frac{S_{yy}(\nu)}{S_{xx}(\nu)}d\nu) - \log(\int S_{xx}(\nu)d\nu)$$
$$\Delta^{\infty}_{Y\to X} = \log(\int S_{xx}(\nu)d\nu) - \log(\int \frac{S_{xx}(\nu)}{S_{yy}(\nu)}d\nu) - \log(\int S_{yy}(\nu)d\nu)$$

Summing up both sides we get:

$$\Delta^{\infty}_{X\to Y} + \Delta^{\infty}_{Y\to X} = -\log(\int |\hat{h}|^2(\nu)d\nu \int \frac{1}{|\hat{h}(\nu)|^2}d\nu) =$$
$$-\log(1 - \mathbb{C}ov(|\hat{h}(\nu)|^2, \frac{1}{|\hat{h}(\nu)|^2}))$$

□

All these theoretical justifications ensures us that our inference method is consistent in a crude sense; that for some simple prior assumptions the method is capable of preferring one direction to other and moreover if the method picks one direction as certain based on the smallness of estimators defined so far, this decision will be well defined.

## 6.3 SIC UNDER NOISE

In this section we derive a preliminary result regarding the effect of additive white noise over our data generating model. Suppose that the data generating model is as follows:

$$Y'_\tau = \mathcal{S}(\{X_t\})(\tau) + N_\tau$$

where $N_\tau$ is white noise with amplitude $\sigma$ and assume that $\{Y_t\}$ is the output of the same system in deterministic condition, i.e. when there is no noise. Then for $\Delta^{\infty}_{X\to Y}$ we have

$$\Delta^{\infty}_{X\to Y'} = \log \frac{\int S_{y'y'}(\nu)d\nu}{\int \frac{S_{y'y'}(\nu)}{S_{xx}(\nu)}d\nu \int S_{xx}(\nu)d\nu} =$$
$$\log \frac{\sigma^2 + \int S_{yy}(\nu)d\nu}{\int \frac{S_{yy}(\nu)}{S_{xx}(\nu)}d\nu \int S_{xx}(\nu)d\nu + \sigma^2 \int S_{xx}(\nu)d\nu \int \frac{1}{S_{xx}(\nu)}d\nu}. \tag{33}$$

Now taking the limit when $\sigma$ tends to infinity one can easily infer that

$$\lim_{\sigma\to\infty} \Delta^{\infty}_{X\to Y'} = \log \frac{1}{\int S_{xx}(\nu)d\nu \int \frac{1}{S_{xx}(\nu)}d\nu}.$$

This means that in the presence of very large noise the sign of the delta expression in forward direction, namely $\Delta^{\infty}_{X\to Y'}$ will be negative except for

the case where the real input to the system is white noise as well. On the other hand for the spectral expression in the backward direction we have

$$\Delta^{\infty}_{Y'\to X} = \log \frac{\int S_{xx}(\nu)d\nu}{\int \frac{S_{xx}(\nu)}{S_{y'y'}(\nu)}d\nu \int S_{y'y'}(\nu)d\nu} = \log \frac{\int S_{xx}(\nu)d\nu}{\int \frac{S_{xx}(\nu)}{S_{yy}(\nu)+\sigma^2}d\nu \left(\sigma^2 + \int S_{yy}(\nu)d\nu\right)}.$$

Once again in the presence of a noise with a very large amplitude one gets:

$$\lim_{\sigma\to\infty} \Delta^{\infty}_{Y'\to X} = \log \frac{\int S_{xx}(\nu)d\nu}{\int S_{xx}(\nu)d\nu} = 0. \tag{34}$$

Equations (33) and (34) indicate that for a white noise with a very large amplitude the spectral expression in forward direction approaches to a negative value and the expression in backward direction approaches to zero; we will see in the next chapter that our causal inference method chooses the causal direction associated with larger spectral expression as the correct causal direction. Therefore in the presence of large noise regime our inference method spoils. We leave a more elaborated analysis of LTIs under noise to later works.

# 7

# ALGORITHMS AND EXPERIMENTS

In this section we define our causal inference algorithm and we apply it to a synthetic data set under deterministic and noisy conditions. We also apply our algorithm to some real world data sets.

## 7.1 SYNTHETIC DATA: COMBINATION OF TWO IIR FILTERS

We have designed an experiment using synthetic data in order to observe the behaviour of estimators in both directions and to get a better idea of designing the decision rule for our causal inference algorithm, which we explain below. The data generating process is as follows. We considered two IIR filters $\mathcal{S}$ and $\mathcal{S}'$, with parameters $(\mathbf{a}, \mathbf{b})$ and $(\mathbf{a}', \mathbf{b}')$ respectively. In each trial $\mathbf{a}$ and $\mathbf{a}'$ are sampled from an isotropic Gaussian distribution with identity covariance matrix, and $\mathbf{b}$ and $\mathbf{b}'$ were sampled from an isotropic multidimensional Gaussian distribution with $0.1$ times the identity covariance matrix. In both cases using rejection sampling we sampled until the coefficients were associate to a stable filter. Also a sequence of length 10000, $\{Z_t\}$ has been sampled by sampling each $Z_t$ from a normal distribution. Then we considered $\{X_t\}$ to be $\mathcal{S}(\{Z_t\})$ and $\{Y_t\}$ to be $\mathcal{S}'(\{X_t\})$. The spectrum for $\{X_t\}$ and $\{Y_t\}$ has been calculated using Welch's method [71]. We repeated this experiment 1000 times. Figure 2 shows an example of the distribution of $\Delta^{\infty}_{X\to Y}$ and $\Delta^{\infty}_{Y\to X}$ where

$$FO(\mathcal{S}) = BO(\mathcal{S}) = FO(\mathcal{S}') = BO(\mathcal{S}') = 5.$$

An important observation to make at this point is that the empirical distribution of estimator in the correct direction is concentrated around zero, however in the wrong direction the estimator stays negative for most of its mass (in this example %97.3). Based on these observation and the theoretical arguments we designed algorithm 1 as our causal inference algorithm. As one can see, the algorithm selects the causal direction associated with the larger spectral expression estimator as the correct causal direction. Interestingly as the following lemma and corollary show, this coincides with choosing the causal direction associated with the spectral expression estimator with the smaller absolute value

**Lemma 10.** *Suppose* $a, b \in \mathbb{R}$ *are given. If* $a + b < 0$ *then* $|a| < |b|$ *iff* $a > b$.

*Proof.* The proof follows by separately considering the cases where $a$ is positive and is non-positive. □

**Corollary 4.** $\Delta^{\infty}_{X\to Y} > \Delta^{\infty}_{X\to Y}$ *iff* $|\Delta^{\infty}_{X\to Y}| < |\Delta^{\infty}_{X\to Y}|$

*Proof.* The proof is a result of lemma 9 and lemma 10. □

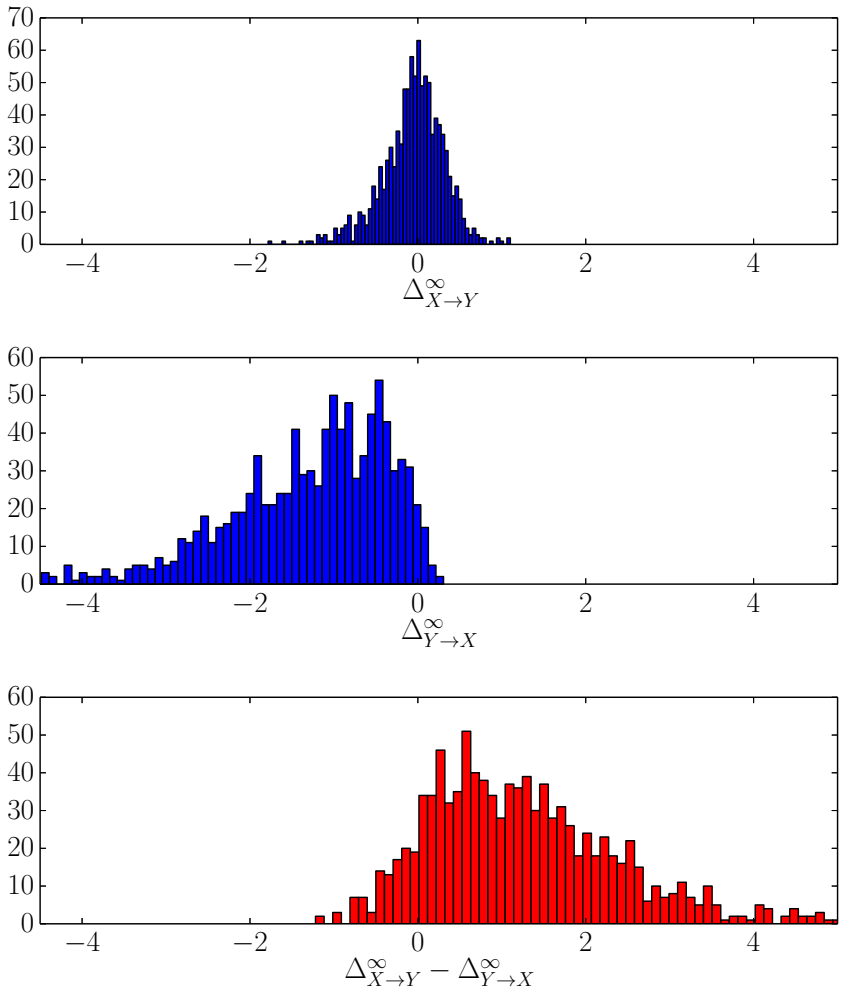



Figure 2: Histogram for the estimators for $\Delta^{\infty}_{X\to Y}$ and $\Delta^{\infty}_{Y\to X}$ in 1000 trials and the difference, $\Delta^{\infty}_{X\to Y} - \Delta^{\infty}_{Y\to X}$ from top to bottom. For more details refer to text

All the confidence intervals in our algorithm are calculated using Wilson's score interval [74]:

$$\frac{1}{1+\frac{1}{n}z^2_{\alpha/2}}\left[\hat{p}+\frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\hat{p}\,(1-\hat{p})+\frac{1}{4n^2}z^2_{\alpha/2}}\right]$$

where $\hat{p}$ is the estimated success rate, $z_{\alpha/2}$ is $100(1-\alpha/2)$-th percentile of the standard normal distribution.

After establishing the inference algorithm, the first test was the effect of change in dimensions of the filter over the performance of the method. In all the examples we set $\alpha = 0.05$ (and therefore $z_{\alpha/2} = 1.96$). Also the parameters for the following simulations are as stated before unless it is explicitly stated otherwise.

**Algorithm 1** SIC_Inference

1: **procedure** SIC_INFERENCE
**Require:** Two time series $\{X_t\}$ and $\{Y_t\}$ are given.
2: $S_{xx} \leftarrow$ spectrum of $X_t$
3: $S_{yy} \leftarrow$ *spectrum of* $Y_t$
4: Calculate $\Delta^{\infty}_{X\to Y}$ and $\Delta^{\infty}_{Y\to X}$ using eqs. (20) and (21)
5: *Inference Step*:
6: **if** $\Delta^{\infty}_{X\to Y} > \Delta^{\infty}_{Y\to X}$ **then return** $X \to Y$
7: **else return** $Y \to X$

### 7.1.1 *The effect of feedforward and feedback orders on performance*

First, we varied the dimension between 2 and 20 and compared the performance of the cases $FO(\mathcal{S}) = BO(\mathcal{S})$ and $BO(\mathcal{S}) = 0$ as can be seen in fig. 3. As one can see with the increase of dimension (when dimension is greater than 3), the performance exceeds %90. It can be seen that the increase of feedback order in the absence of feedforward order results in a drop in performance, however in the presence of feedforward order equal to feedback order, the performance does not diminish.

Next we did a similar comparison between $FO(\mathcal{S}) = BO(\mathcal{S})$ and $FO(\mathcal{S}) = 0$, which can be seen fig. 4. In this case one can realise that the feedback order of an IIR is the reason of the inferior performance in for the filter with zero FO in these experiments.

### 7.1.2 *The effect of additive noise on performance*

We also used our inference method on a variant of the previous data generating mechanism where an additive zero mean Gaussian noise with different standard deviations $\sigma$ has been added to the output of the filter. The plot for the performance of the method in two different cases where $FO = BO = 5$ and where $FO = 0$ and $BO = 5$ can be seen in fig. 5. When FO is large the performance of the method is quite robust to noise but when it is zero the performance drastically decreases even in low noise regime scenarios. Nevertheless in both cases the performance drops by increasing the amplitude of the noise which is in line with the theoretical results derived in section 6.3

## 7.2 REAL WORLD EXAMPLES

Since the aim of these thesis was to justify the applicability of this method over real data, we have applied our method over a few different examples of real data where the ground truth about the causal structure of the data is known a priori.

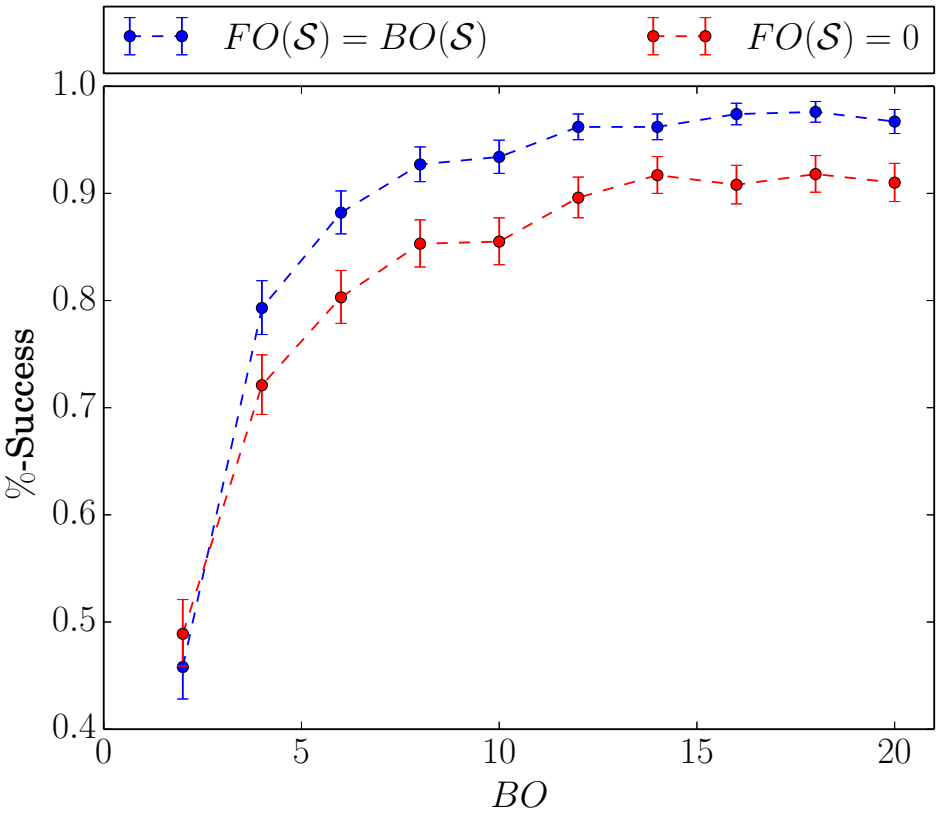



Figure 3: Comparison of performance of the inference algorithm in deterministic (no noise) case, where feedback order is varying and feedforward order (red plot) is either zero (blue plot) or equal to feedback order.

### 7.2.1 *Gas Furnace*

We have applied our framework on gas furnace data available in [11, pp. 548]. The data is the consumed gas rate $\{X_t\}$ by a gas furnace and the produced rate of $CO_2$, $\{Y_t\}$. The ground truth is assumed to be $\{X_t\} \to \{Y_t\}$. Since the data length in this data set is quite small (296 data points), and the result of our method is very sensitive to the estimation of power spectral density we have applied our method with taking different lengths of window sizes for Welch method into account. The results has been calculated as the difference between $\Delta^{\infty}_{X\to Y} - \Delta^{\infty}_{Y\to X}$ and plotted as a function of window length which was ranging from 50 to 149 (inclusive) and the plot can be seen in section 7.2.2. The upper bound of window lengths has been chosen in a way so that the variance of power spectral density estimation with Welch method could be reduced by 2, comparing to the case where the full time window is used for spectral density estimation; this is done because our estimators are highly dependent on values of the spectrum close to zero and this variance reduction can help us to prevent large estimation error for such points. As can be seen this difference is always positive and our method is able to correctly infer the right causal direction. TiMiNO and Granger causality are also both capable of inferring the correct direction in this case [48].

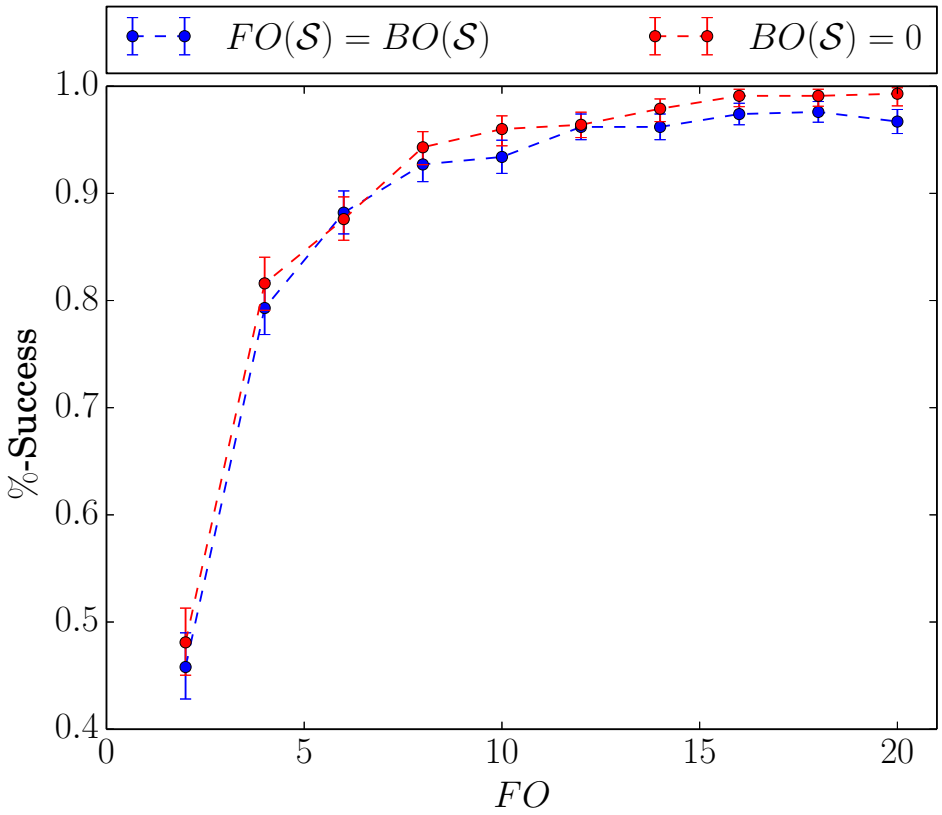



Figure 4: Comparison of performance of the inference algorithm in no noise case, where feedback order is varying and feedback order (red plot) is either zero (blue plot) or equal to feedforward order.

### 7.2.2 *Old Faithful Geyser*

We have applied our method to data recorded from Old Faithful Geyser in Yellow Stone National Park, Wyoming, USA [8]. The data recorded is the time interval between successive eruptions (taking it as $\{X_t\}$) and duration of the subsequent eruption (represented here as $\{Y_t\}$. Since part of the data has been gathered during the night, some of the $X_t$'s are only reported as short, medium or long intervals. Following the analysis in [8] we replace these values by 2, 3 and 4 minutes. We will consider $X \to Y$ as the ground truth causal structure as has been done in [48]. Since the data set size is small (298), following the argument of the previous example, we plot the difference $\Delta^{\infty}_{X\to Y} - \Delta^{\infty}_{Y\to X}$ as a function of the chosen window length which ranges from 50 to 149 (inclusive) and the results can be found in section 7.2.2. As can be seen, our method chooses the correct causal direction for all of the window lengths in the interval. It has been reported in [48], TiMiNO is also capable of inferring the correct causal structure however both linear and nonlinear Granger causality methods fail on the problem.

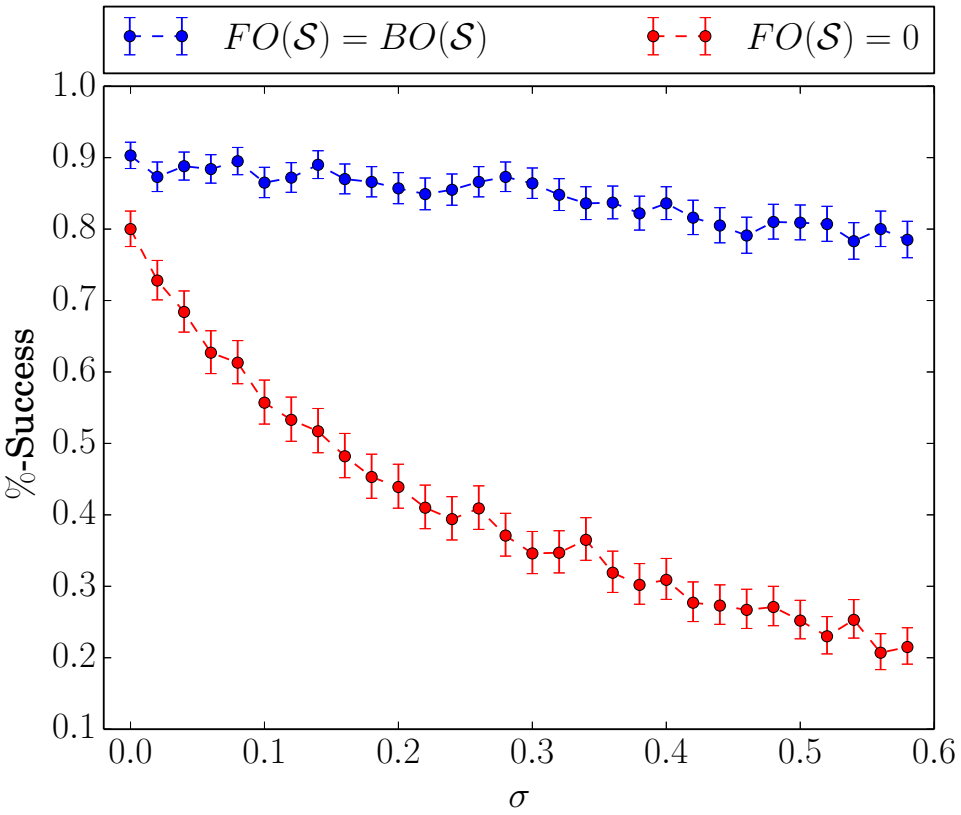
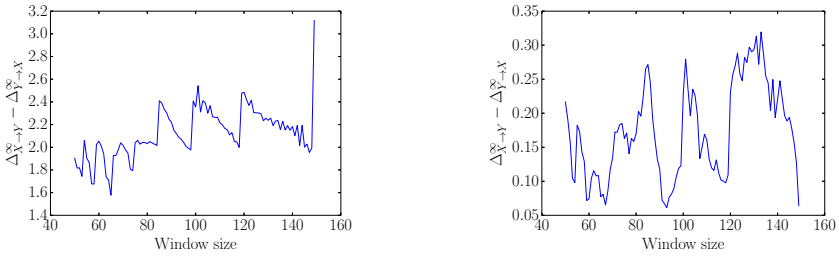



Figure 5: Comparison of performance of the inference algorithm in case where noise with different amplitudes is to data. Two cases are considered. For more details see the text.

### 7.2.3 *Neural Data: LFP recordings of the Rat Hippocampus*

It is known that contrary to neocortex where the connections between different regions is usually bidirectional, the connections between different parts of hippocampal formation are mostly unidirectional [5, pp. 38]. An important example of this type is the unidirectional connection between areas CA3 and CA1 through Schaffer collaterals [5, pp. 38]. Despite this anatomical fact the only study of causality based on local field potential recordings of CA1 and CA3 of the hippocampus of the rat during sleep - to the best knowledge of the authors - reports [9] that Granger causality infers strong bidirectional relations between the two areas. [9] explains the possible reasons of this results as the long-loop feedback involving cortex and medial septum, and also diffuse connections from CA1 to CA3.

To do a comparison with Granger causality over neural data we applied our framework on the recordings of the same sites using a different data set that is publicly available[1]. The local field potential has been recorded using an 8 shank probe having 64 channels downsampled to 1250Hz and the shanks has been divided equally between CA1 and CA3 areas (leaving 32 channels for each area) which the recorded voltage has been amplified by 1000. For more information on the details of gathered data please refer to




Gas furnace Old Faithful Geyser

Figure 6: The plots for difference between the estimators of spectral expressions in both directions as a function of window length chosen for Welch method. The plot for gas furnace is on the left and for old geyser is on the right. As one can see algorithm 1 will always pick the correct causal direction independent of the window size.

[1]. The information used belongs to the rat named "vvp01" and for two different sessions of linear walking (where the rat had to walk a straight path) and sleeping that are named as "2006-4-9_17-29-30" and "2006-4-9_18-43-47" respectively for the first three minutes of recording. The Granger causality method that we used was based on Vector Auto Regressive model fitting and the comparison of Residual Sum of Squares (RSS) between the outcome of regressing LFP of CA3 to LFP of CA1 and vice versa. For this purpose we used an available implementation of Granger causality by statsmodel [3] for Python programming language; In this implementation when checking whether $\{X_t\}$ is the cause of $\{Y_t\}$ the null hypothesis is true, if $\{X_t\}$ does NOT Granger causes $\{Y_t\}$. Since SIC assumes a prior that $\{X_t\} \to \{Y_t\}$ $\{Y_t\} \to \{X_t\}$ we considered a forced decision scheme for Granger causality, i.e. for any Granger causality test between two time series we select the one with lower p-value as the correct causal direction excluding the possibility of non of them being the cause of the other. Following the usual methodology of causality analysis [9, 15] we have divided the duration of three minutes to 180 intervals of one second to reduce the effect of nonstationarity in data analysis. Since there were 32 channels available for each site we have calculated the success performance of the method in a time window as the percentage of the channel pairs of CA1 to CA3 which are in total 1024 for any time window of one second.

The results for the linear session can be seen in fig. 7. The performance of our inference method outperforms the performance of linear Granger causality with a great difference. Moreover it stays (except for one time window) above the chance level discrimination line (the confidence intervals included), which is in line with the unidirectional anatomical projections from CA3 area to CA1.

The results for the sleeping session are indicated in fig. 8. Despite the poor performance of our method in this case we believe the results are highly

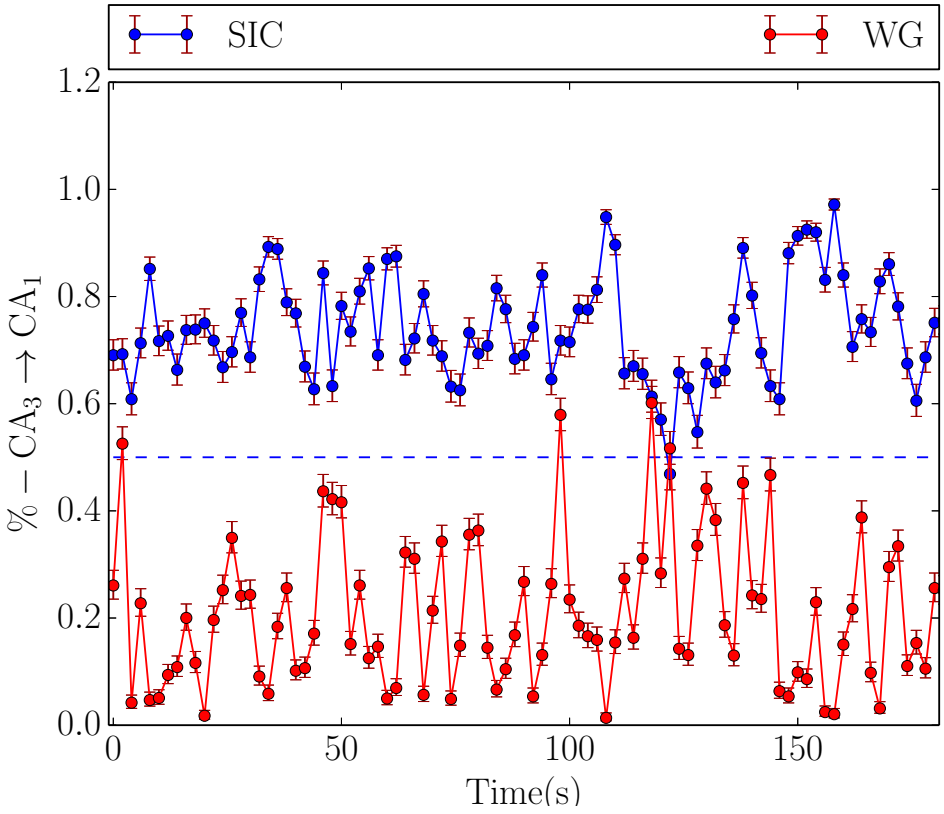



Figure 7: Comparison of performance of the linear Granger causality and spectral independence methods during the linear session for the mice "vvp01". The dashed line indicates when the performance is equal to fifty percent. For more information please refer to text.

promising; First of all even in this case the method still outperforms linear Granger causality in inferring, the causal direction which is in line with anatomical unidirectional connection. Moreover it stays above the chance level discrimination for most of the time windows. Also our preliminary analysis suggests that the reason for the poor performance in this case is relevant to nonstationarity of the recorded LFP in both areas during those specific time windows. For example for the time window starting at second 106s (which the method gives close to zero performance), the shape of the recordings from all the channels in both recording sites can be seen in fig. 9 (upper figure) and fig. 9 (lower figure). One can appreciate that in both cases the time series lacks the appearance of a weakly stationary times series; which can be acknowledged by comparing these time windows to the neighbouring ones, especially the ones for which the performance of the algorithm is relevantly high (also by comparing the activity in these areas during the linear session).

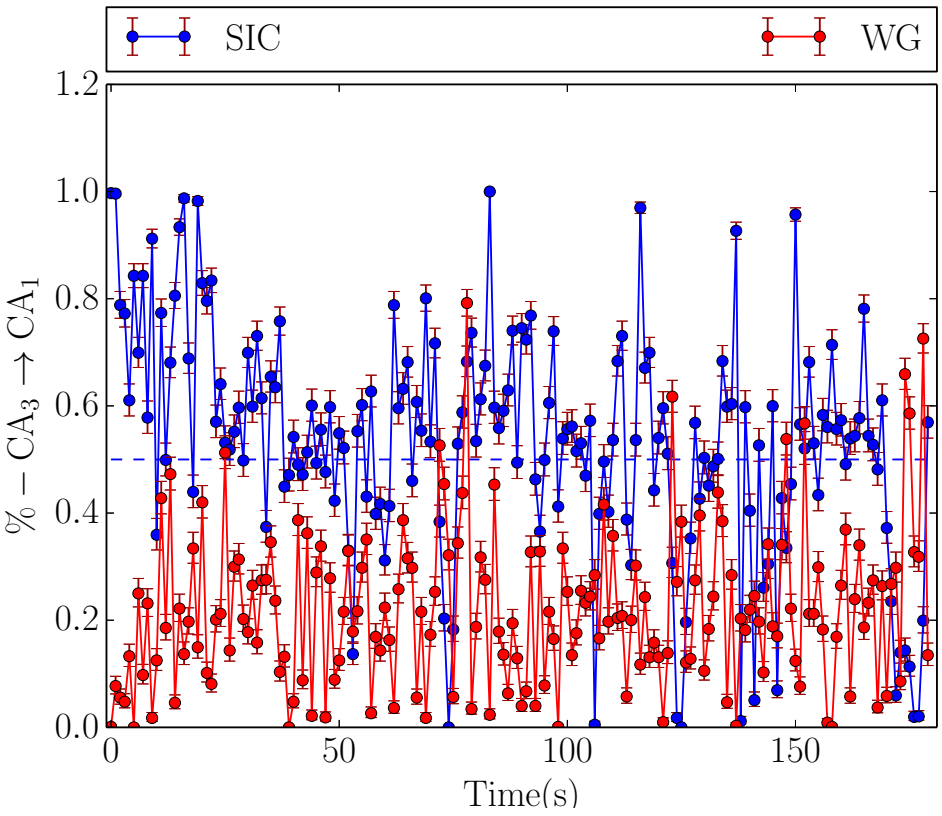



Figure 8: Comparison of performance of the linear Granger causality and spectral independence method in the sleeping session for the mice "vvpo1". The dashed line indicates when the performance is equal to fifty percent. For more information please refer to text.

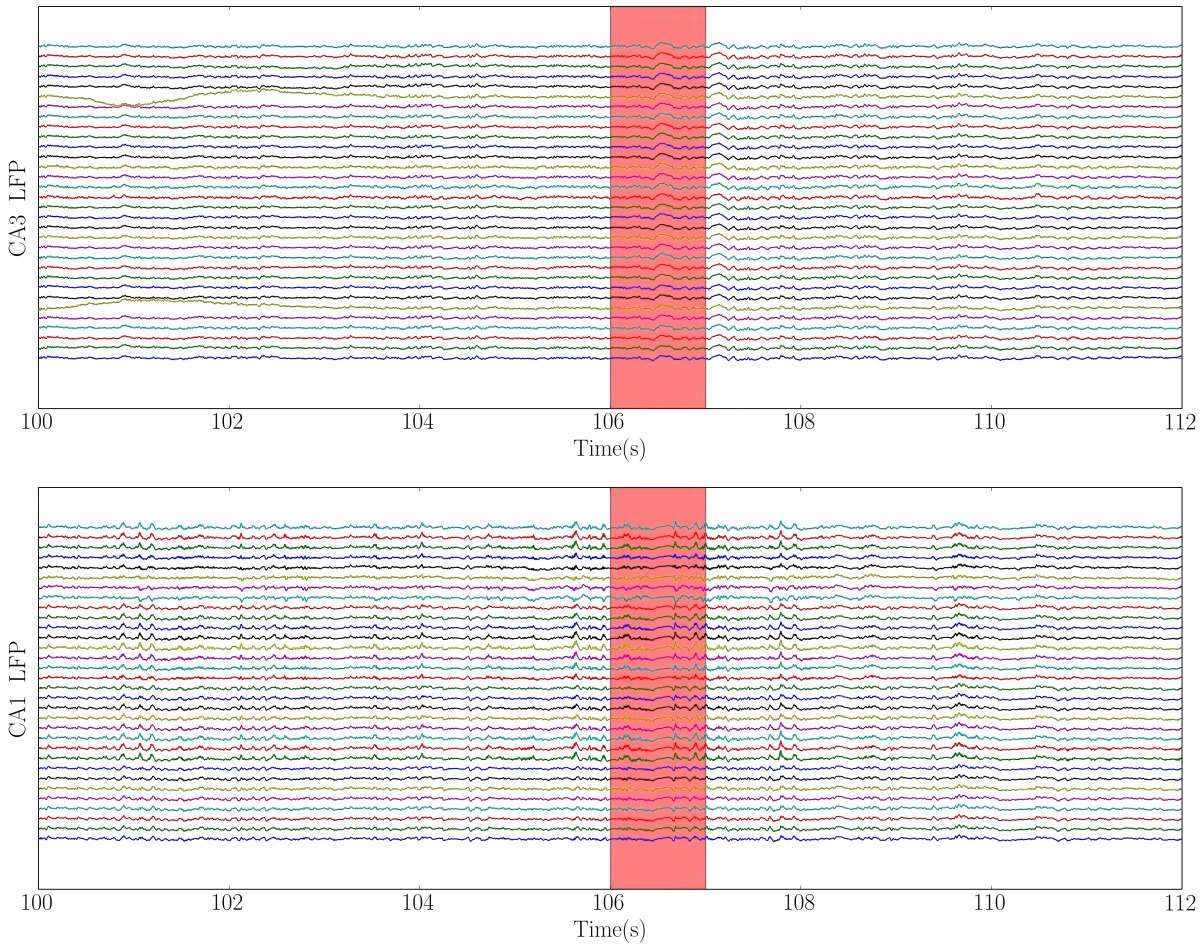



Figure 9: LFP recordings of all the channels for period between 100s and 112s. The above figure is the LFP from 32 channels of CA3 area. Similarly in the bottom plot we have presented the LFP recordings of 32 channels of CA1 area for rat "vvp01" during sleep. The red window correspond to a 1s time window (at 106s) where SIC fails strikingly. One can appreciate that the signal is nonstationary in this time window, both in CA1 and CA3.

# 8

# DISCUSSION AND OUTLOOK

In this work, we have introduced a method of causal inference, to infer the causal direction between a pair of time series where only one of them is causally influencing the other. This method is based on a recently developed framework [33] that infers the causal direction between observed pairs of observations by exploiting the fact that the mechanism that generates the effect from the cause is in some way independent from the distribution of the cause.

We have showed that this method is a modification of a causal inference method for high dimensional linear relationships when the dimension tends to infinity, that incorporates the temporal structure of observations. We have derived some identifiability results based on concentration of measure phenomenon for the cases that the data generating process is an FIR filter over discrete and circular weakly stationary processes (c.f. to theorems 11 and 12). Moreover we proved that our main expression for deriving the causal direction, the "spectral estimator", cannot be arbitrarily small in both directions (c.f. to lemma 9). We also derived some preliminary results on the performance of the method under additive noise (c.f. to section 6.3).

Based on some simulations over a toy model and our identifiability results we have developed our causal inference algorithm (c.f. to algorithm 1). We showed that the algorithm is effective on synthetic data, in the deterministic case and also in the nondeterministic case under a weakly noisy regime. We also successfully applied our method to real data; especially we showed that our method outperforms Granger causality on inferring the correct causal structure between the electrical activities recorded from CA1 and CA3 hippocampal areas of rat hippocampus which has been anatomically verified to be a unidirectional connection. This, along the other examples on real world as well as synthetic data already indicates that inferring causal relations on time series might as well be exploited by causal inference methods based on postulates different than what has been proposed by Wiener and Granger.

## 8.1 SHORTCOMINGS AND FUTURE GOALS

Although our method performed well in experiments that have been done so far, its poor performance (despite its superiority to Granger causality) over neural data recorded from hippocampal formation in sleeping rat (see section 7.2.3) shows that the method can fail, specially when the assumptions of our method are not met; this can be because of nonstionarity of data and/or because the modelling assumption does not hold, i.e. the real

functional relationship between the two time series cannot be modelled effectively using LTI systems. Both of these scenarios can easily be realized in real world problems. Therefore applying some approximation techniques (like the small window sizes to assure stationarity that has been applied in section 7.2.3) can help us better model the behaviour of a real world phenomenon and therefore to derive more accurate results. Besides, linear models are sometimes unable to capture all the details of a nonlinear system. Developing a causal inference method based on ICM that could directly address the problem of causal inference on time series that are causally related through some nonlinear dynamics might significantly improve the performance of the method when the nonlinear dynamics are complex enough; as previously mentioned the same idea has been exploited to address the causal inference problem for static observables (see section 3.2.1).

Another important issue was that our comparison analysis between SIC and Granger causality was elementary since we haven't addressed the problem of developing a statistical test for SIC method yet; an availability of p-values for our method and a proper statistical test for it can greatly improve the accuracy of the comparisons between this method and other causal inference method for time series. One way to realize this (as been suggested in [33] for a similar problem) is to use regression methods to fit an FIR filter to data and to calculate p-values by applying rotation matrices over coefficients of this FIR filter and recalculating the estimator values. Since FIR filters are a very small family of LTI filters, a more advanced method needs to be developed for statistical test, maybe based on the same idea but by fitting ARMA filters and/or by stronger identifiability results that could help to calculate p-values for statistical tests.

Our preliminary works indicates that there might exist stronger ties between ICM based method for nonlinear data and the SIC method; one needs to realize that the causal inference method developed in [20] is based on the idea that $p_X$ or the density of cause is independent from the function $f$ that relates it to the effect $Y$. This independence can indeed be phrased in terms of independence between stochastic processes that $f$ and $p_X$ are sampled from. We leave the possibility of deriving such a connection between these two methods to our future works.

Finally the identifiably results under noise were preliminary. We believe that one can derive more concise results relevant to the behaviour of this causal inference method under additive white noise. We leave this extension to our future works.

Part I

# APPENDIX

# A

# APPENDIX

In this appendix we develop the SIC method with a different approach and by means of linear operator theory.

## A.1 SIC IN LINEAR OPERATOR THEORY

Suppose we want to generalize the causal inference framework for deterministic linear relationship introduced in [34, 77] to infinite dimensional case where the input and output of the linear mapping are weakly stationary stochastic processes. In the following we treat the continuous case (The inferences can be derived for discrete case just by replacing integral with summation). As such assume the linear relationship

$$\{Y_t\} = \mathcal{L}(\{X_t\}),$$

(defined below), where $\{X_t\}$ and $\{Y_t\}$ are zero mean continuous Gaussian processes with covariance functions $\mathbf{K}_{xx}$ and $\mathbf{K}_{yy}$. With some overload of nation we represent $\{X_t\}$ and $\{Y_t\}$ as $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ respectively. Suppose $\mathcal{L}$ maps $\mathbf{X}$ to $\mathbf{Y}$ with the following integral transform

$$\mathbf{Y}(y) = \int \mathcal{L}(x, y)\mathbf{X}(x)dx.$$

Then its known from the literature [53] that

$$\mathbf{K}_{yy} = \mathcal{L}\mathbf{K}_{xx}\mathcal{L}^\top.$$

If moreover $\mathcal{L}$ and $\mathbf{K}_{xx}$ would be time invariant we can write:

$$\begin{aligned} \Phi_{\mathbf{K}_{yy}} &= \Phi_{\mathcal{L}} * \Phi_{\mathbf{K}_{xx}} * \Phi_{\mathcal{L}^\top} \\ \Phi_{\mathcal{L}\mathcal{L}^\top} &= \Phi_{\mathcal{L}} * \Phi_{\mathcal{L}^\top} \end{aligned} \tag{35}$$

and based on lemma 3, $\mathbf{K}_{yy}$ and $\mathcal{L}\mathcal{L}^\top$ are translation invariant. Since $\mathcal{L}\mathcal{L}^\top$ is positive definite it follows that $\Phi_{\mathcal{L}} * \Phi_{\mathcal{L}^\top}$ is a positive definite function. Based on Bochner's theorem and corollary 1 we get

$$\begin{aligned} \mathcal{T}_{\mathcal{L}\mathcal{L}^\top} = \Phi_{\mathcal{L}\mathcal{L}^\top}(0) = \int |\Phi_{\mathcal{L}}|^2 = \int S_{\mathcal{L}\mathcal{L}^\top} \\ \mathcal{T}_{\mathbf{K}_{xx}} = C_X(0) = \int S_{xx} \\ \mathcal{T}_{\mathbf{K}_{yy}} = C_Y(0) = \int S_{yy} \end{aligned}$$

where $S_{\mathcal{L}\mathcal{L}^\top}$ is the spectral density for $\phi_{\mathcal{L}\mathcal{L}^\top}$. Therefore trace condition in this case reads:

$$\int S_{yy} = \int S_{\mathcal{L}\mathcal{L}^\top} \int S_{xx}$$

If we apply Fourier transform on both sides of eq. (35) we get:

$$S_{yy}(\nu) = S_{\mathcal{L}\mathcal{L}^\top}(\nu) S_{xx}(\nu)$$

Which simplifies the trace condition to:

$$\int S_{yy} = \int \frac{S_{yy}}{S_{xx}} \int S_{xx}$$

Therefore we can rephrase our spectral independence criterion in the context of linear operator theory as follows:

**Postulate 4. (SIC in Linear Operator Theory)** *Let* $f, g$ *and* $h$ *to be positive definite functions on* $\mathbb{R}$ *and suppose* $g = f * h$ *and* $\mu_f, \mu_g$ *and* $\mu_h$ *their Fourier transform measures respectively. Moreover assume these measures are absolutely continuous with respect to a reference measure with spectral densities* $S_f, S_g$ *and* $S_h$*. We say* $f$ *and* $h$ *are chosen independently if* $\mathrm{Cov}(S_f, \frac{S_g}{S_f}) = 0$

# BIBLIOGRAPHY


[1] CRCNS - Collaborative Research in Computational Neuroscience, hc3 data set., . URL http://crcns.org/data-sets/hc/hc-3/about-hc-3.

[2] A Chronological History of Causation for Environmental Scientists, . URL http://www.epa.gov/caddis/si_history.html.

[3] Statsmodels: Statistical Analysis Toolkit for Python, . URL http://statsmodels.sourceforge.net/.

[4] Jorge Renner Cardoso de Almeida, Amelia Versace, Andrea Mechelli, Stefanie Hassel, Karina Quevedo, David Jerome Kupfer, and Mary Louise Phillips. Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biological psychiatry*, 66(5):451–459, 2009.

[5] Per Andersen, Richard Morris, David Amaral, Tim Bliss, and John O'Keefe. *The hippocampus book*. Oxford University Press, 2006.

[6] J Arnhold, P Grassberger, K Lehnertz, and CE Elger. A robust method for detecting interdependences: application to intracranially recorded eeg. *Physica D: Nonlinear Phenomena*, 134(4):419–430, 1999.

[7] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis*, volume 39. Springer, 2005.

[8] A Azzalini and AW Bowman. A look at some data on the old faithful geyser. *Applied Statistics*, pages 357–365, 1990.

[9] LA Baccala, K Sameshima, G Ballester, AC Do Valle, and C Timo-Iaria. Studying the interaction between brain structures via directed coherence and granger causality. *Applied Signal Processing*, 5(1):40, 1998.

[10] Lé Bottou. From machine learning to machine reasoning: an essay. *Machine Learning*, 94:133–149, January 2014. URL http://leon.bottou.org/papers/bottou-mlj-2013.

[11] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

[12] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer, 2009.

[13] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[14] György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012.

[15] Alex J Cadotte, Thomas B DeMarse, Thomas H Mareci, Mansi B Parekh, Sachin S Talathi, Dong-Uk Hwang, William L Ditto, Mingzhou Ding, and Paul R Carney. Granger causality relationships between local field potentials in an animal model of temporal lobe epilepsy. *Journal of neuroscience methods*, 189(1):121–129, 2010.

[16] Gregory J Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM (JACM)*, 22(3):329–340, 1975.

[17] C Chatfield. The analysis of time series: An introduction, 5* edition.

[18] Mario Chávez, Jacques Martinerie, and Michel Le Van Quyen. Statistical assessment of nonlinear causality: application to epileptic eeg signals. *Journal of Neuroscience Methods*, 124(2):113–128, 2003.

[19] Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *The Journal of Machine Learning Research*, 9:967–991, 2008.

[20] P. Daniusis, D. Janzing, K. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 143–150, Corvallis, Oregon, 2010. AUAI Press.

[21] Mingzhou Ding, Yonghong Chen, and Steven L Bressler. 17 granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*.

[22] FR Drepper. Asymptotically stable phase synchronization revealed by autoregressive circle maps. *Physical Review E*, 62(5):6376, 2000.

[23] Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[24] William A Gardner, Antonio Napolitano, and Luigi Paura. Cyclostationarity: Half a century of research. *Signal processing*, 86(4):639–697, 2006.

[25] Claude Gasquet and Patrick Witomski. *Fourier analysis and applications: filtering, numerical computation, wavelets*, volume 30. Springer, 1999.

[26] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[27] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[28] Robert M Gray. *Toeplitz and circulant matrices: A review*. now publishers inc, 2006.

[29] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*, volume 2. Oxford Univ Press, 1992.

[30] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jan R Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[31] Aapo Hyvärinen, Shohei Shimizu, and Patrik O Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the 25th international conference on Machine learning*, pages 424–431. ACM, 2008.

[32] Shunsuke Ihara. *Information theory for continuous systems*, volume 2. World Scientific, 1993.

[33] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *Information Theory, IEEE Transactions on*, 56(10):5168–5194, 2010.

[34] Dominik Janzing, Patrik O. Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 479–486, Haifa, Israel, June 2010. Omnipress. URL http://www.icml2010.org/papers/576.pdf.

[35] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[36] Dominik Janzing, Bastian Steudel, Naji Shajarisales, and Bernhard Schölkopf. Justifying information-geometric causal inference. *arXiv preprint arXiv:1402.2499*, 2014.

[37] GM WATTS JENKINS and S Watts. Dg (1968): Spectral analysis and its applications.

[38] Kitti Kaiboriboon, Hans O Lüders, Mehdi Hamaneh, John Turnbull, and Samden D Lhatoo. Eeg source imaging in epilepsy, practicalities and pitfalls. *Nature Reviews Neurology*, 8(9):498–507, 2012.

[39] Achim Klenke. *Probability theory: a comprehensive course*. Springer, 2007.

[40] Andrei Nikolaevich Kolmogorov. Three approaches to the quantitative definition of information*. *International Journal of Computer Mathematics*, 2(1-4):157–168, 1968.

[41] Michel Le Van Quyen, Claude Adam, Michel Baulac, Jacques Martinerie, and Francisco J Varela. Nonlinear interdependencies of eeg signals in human intracranially recorded temporal lobe seizures. *Brain research*, 792(1):24–40, 1998.

[42] Jan Lemeire and Erik Dirkx. Causal models as minimal descriptions of multivariate systems.

[43] Gabriele Lohmann, Kerstin Erfurth, Karsten Müller, and Robert Turner. Critical comments on dynamic causal modelling. *Neuroimage*, 59(3): 2322–2329, 2012.

[44] Anthony Randal McIntosh. Towards a network theory of cognition. *Neural Networks*, 13(8):861–870, 2000.

[45] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[46] Donald B Percival. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[47] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2436–2450, 2011.

[48] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using structural equation models. *arXiv preprint arXiv:1207.5136*, 2012.

[49] Giovanni Pistone, Maria Piera Rogantin, et al. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, 1999.

[50] R Quian Quiroga, J Arnhold, and P Grassberger. Learning driver-response relationships from synchronization patterns. *Physical Review E*, 61(5):5142, 2000.

[51] Michel Le Van Quyen, Jacques Martinerie, Claude Adam, and Francisco J Varela. Nonlinear analyses of interictal eeg map the brain interdependences in human focal epilepsy. *Physica D: Nonlinear Phenomena*, 127(3):250–266, 1999.

[52] Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *Neuroimage*, 49(2):1545–1558, 2010.

[53] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[54] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

[55] Mario Rosanova, Olivia Gosseries, Silvia Casarotto, Mélanie Boly, Adenauer G Casali, Marie-Aurélie Bruno, Maurizio Mariotti, Pierre Boveroux, Giulio Tononi, Steven Laureys, et al. Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain*, page awr340, 2012.

[56] Walter Rudin. *Fourier analysis on groups*. John Wiley & Sons, 2011.

[57] YOICHI Saito and HIROSHI Harashima. Tracking of information within multichannel {EEG} record causal analysis in {EEG}. *Yamaguchi N, Fujisawa K (eds) Recent advances in {EEG} and {EMG} data processing. Elsevier*, pages 133–146, 1981.

[58] 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, 1996.

[59] SM Schnider, RH Kwong, FA Lenz, and HC Kwan. Detection of feedback in the central nervous system using system identification techniques. *Biological cybernetics*, 60(3):203–212, 1989.

[60] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[61] D. Serre. *Matrices: Theory and Applications*. Graduate Texts in Mathematics. Springer, 2010. ISBN 9781441976833. URL http://books.google.de/books?id=pKKrFnINccIC.

[62] Cosma Rohilla Shalizi. Methods and techniques of complex systems science: An overview. In *Complex systems science in biomedicine*, pages 33–114. Springer, 2006.

[63] Herbert A Simon. Spurious correlation: A causal interpretation*. *Journal of the American Statistical Association*, 49(267):467–479, 1954.

[64] Herbert A Simon. *On the definition of the causal relation*. Springer, 1977.

[65] Herbert A Simon and Nicholas Rescher. Cause and counterfactual. *Philosophy of Science*, pages 323–340, 1966.

[66] Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

[67] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[68] Klaas Enno Stephan and Alard Roebroeck. A short history of causal modeling of fmri data. *Neuroimage*, 62(2):856–863, 2012.

[69] Pedro A Valdes-Sosa, Alard Roebroeck, Jean Daunizeau, and Karl Friston. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.

[70] 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, 2011.

[71] Peter D Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2): 70–73, 1967.

[72] Norbert Wiener. The theory of prediction, 1956.

[73] Jon Williamson. Causality. In *Handbook of philosophical logic*, pages 95–126. Springer, 2007.

[74] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158): 209–212, 1927.

[75] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[76] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

[77] Jakob Zscheischler, Dominik Janzing, and Kun Zhang. Testing whether linear equations are causal: A free probability theory approach. *CoRR*, abs/1202.3779, 2012.

## COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

*Final Version* as of September 16, 2014 (`classicthesis` version 4.0).