MUTUAL INFORMATION BASED MEASURES ON COMPLEX INTERDEPENDENT NETWORKS OF NEURO DATA SETS

A thesis presented for the degree of

Doctor of Philosophy in Mathematics of Imperial College London

and the

Diploma of Imperial College

by

FATIMAH ABDUL RAZAK

supervised by PROF. HENRIK JELDTOFT JENSEN

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

MARCH 2013

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed:

Copyright

Abstract

We assume that even the simplest model of the brain is nonlinear and ‘causal’. Proceeding with the first assumption, we need a measure that is able to capture nonlinearity and hence Mutual Information whose variants includes Transfer Entropy is chosen. The second assumption of ‘causality’ is defined in relation to prediction ala Granger causality. Both these assumptions led us to Transfer Entropy. We take the simplest case of Transfer Entropy, redefine it for our purposes of detecting causal lag and proceed with a systematic investigation of this value. We start off with the Ising model and then moved on to created an amended Ising model where we attempted to replicate ‘causality’. We do the same for a toy model that can be calculated analytically and thus simulations can be compared to its theoretical value. Lastly, we tackle a very interesting EEG data set where Transfer Entropy shall be used on different frequency bands to display possible emergent property of ‘causality’ and detect possible candidates for causal lag on the data sets.

Acknowledgement

Praise be to Allah for this enlightening journey which has brought me to many different places, tested my limits and hopefully made me a better person. My experience in Imperial have been a pleasant one mainly due to the people I have met. First and foremost to my esteemed supervisor, Professor Henrik Jeldtoft Jensen who has time and time again guided me in the right direction with his great insights. I really appreciate the fact that he always manages to motivate me and that he has given me the freedom to explore. The very supportive Complexity and Networks group made me feel right at home here at Imperial. Fellow PhD students Miss Nicky Nicola Zachariou, Dr. Paul Expert and Dr. Giovanni Petri were the individuals most responsible for this welcoming atmosphere. Miss Zachariou in particular deserves special mention. Not only did she proof read this thesis but she is also the unofficial secretary of the group and is usually the one responsible for bringing us together. It is no surprise that she 'shakes' the earth wherever she goes.

The words of wisdom from the more experienced member (and former members) of the group namely Prof Kim Christensen, Dr Renauld Lambiotte, Dr Elsa Arcaute and Dr Micheal Gastner were most helpful. The enjoyable discussions with newer members of the group Xiaogeng Wan, Kristijonas Broga, Kishan Manani, Duccio Piovani, Alvis Tang, Dr. Roseli Wedemann and Shama Rahman, are always stimulating. The communication with the Mathematics Department of Imperial College facilitated the PhD greatly, many thanks to Svanide Rusudan and Prof John Gibbons. The funding for my PhD was provided by the Malaysian government in conjunction with Universiti Kebangsaan Malaysia (UKM) a.k.a National University of Malaysia. I would like to extend my gratitude to the School of Mathematical Sciences and Department of Mathematics as well as the Human Resource department of UKM for being so helpful and understanding.

The Malaysian London community have been essential in making me feel at home in London by providing avenues for festive celebrations and spiritual sessions. The conducive environment provided by the lovely people of The Victoria League of Commonwealth Friendship and MARA Hostel contributed greatly towards development of ideas for the PhD. Special thanks to Latifah Karim and Ummu Tasnim Husin for proof reading my thesis and always being there to listen to my problems. They form part of a caring circle of friends that are always curious about my wellbeing and ever ready to lend a hand. This circle of friends is also comprised of but not limited to Nurul Hana Mokhtar, Deena Yacob and husband Nazri Nawi, Syahida Zainuddin, Ahmad Mahfuz Ghazali, Raja Hafida, Nur Izzah Bakar and Siti Mariam Shafie.

Last but certainly not least, a big thank you to my family in Malaysia for the unconditional love and never ending prayers on my behalf. My father Abdul Razak Abu Bakar for his soothing reassurances and my mother Sallina Mohd Zain for her heartwarming e-mails. My brothers and sister for being the nicest siblings one could ever have (despite occasional naughtiness). I have to apologize for not calling often enough. I am truly indebted to all of you. I dedicate this thesis to you my family and friends.

Table of contents

Chapter 1

The brain as a complex system

Statistical mechanics works towards understanding the macroscopic behaviour of a system from the microscopic interaction of its part. Lately the field has taken interest in complex systems, loosely defined as a system where the components self-organize into a critical state [6, 27, 53]. Criticality (the critical state) occurs when local distortions propagates throughout the entire system and observables are scale free [27, 53]. In this introductory chapter, we will first explain how the brain is essentially a complex system and how it is logical for it to operate near criticality. Then we look at the methods of analysing neuro data sets and why EEG data sets suits our purposes. After looking at some of the many different measures people have used on EEG data sets, we will most importantly explain our simplistic view of the brain and why we are looking for a measure that should essentially be nonlinear and be able to detect possible 'causality'.

1.1 Complexity and the brain

Although there is no universally agreed upon definition of complex system, there are a few criteria that most will say define the system, mainly that it is composed of a large number of interacting components that give rise to emergent hierarchical structures and that the components typically change with time [54]. There is increasing belief among neurologists is that the brain is complex [20, 43, 92]. In fact, it has been suggested that this wonderful brain of ours, could possibly be the most complex system of them all, due to its capability to

represent the complex outer world to us humans [6]. Indeed, we are what we are perceived to be. A simple way to view the brain as a complex system is to imagine the neurons as individual components and the brain as a whole that emerges from cooperation of the neurons.

1.1.1 Complex systems and criticality

In other words, the brain is viewed as a system that exhibits complex behaviour where hundred billions of neurons self organize to function as one entity. However, some neurons seem to work together with one another more than others depending on the performed function [24, 92]. Consequently, it is believed that the brain can be divided into its functional parts [25, 36]. Therefore, one could also consider the functional parts (encompassing certain areas of the brain) as the components that cooperate to become the brain as a whole.

This is where a contradiction arises, on one hand the brain needs to be segregated for it to specialize and efficiently respond to specific stimulus (or lack of it) but on the other hand scientist have confirmed from many different sources and observations that virtually all perceptual or cognitive tasks are the end result of large scale and distributed activities, often times involving spatially disconnected area [25, 24]. This integration versus segregation issue is succinctly summarized in [92]. This apparent contradiction can be explained beautifully if we look at the brain as complex system. In order to further understand this currently prevalent paradigm about how the brain works we need to first understand criticality.

Criticality is the emergence of the components to work together as one. A familiar example will be the phase transition occurring when water evaporates into vapor. A phase transition with long range interactions (diverging correlation lengths) is critical. Basically in a complex system, criticality occurs when the system appears to act cooperatively as the result of individuals interactions [27, 53]. Therefore as a complex system, the brain is completely connected (acting cooperatively) and integrated at criticality as well as being segregated (acting individually or according to its functional speciality).

If a system is critical, there will be long range correlations as well as short range ones; in fact all scales of correlation should be present. Thus, theoretically the system will appear

identical no matter the scale at which we probe it, hence the term “scale free”. This self similar and scale free behaviour is fractal in nature which leads to the belief that a complex system is also fractal [27, 53]. This fact allows us to use scale free and self similarity as indicators of a system being in a critical state. The ability of certain data sets to appear self similar regardless of hierarchies and coarse graining is often taken to be an indication of criticality. However caution must be taken here since there can be other causes [104] for a system to be fractal.

1.1.2 The Ising model at phase transition

A system is critical in a sense that all members influence each other. In a complex system the critical state is established solely because of individual interactions [6, 27, 53] but an equilibrium system needs to be fine tuned to obtain this criticality. Perhaps the simplest example of an equilibrium system is the Ising model [27]. It displays emergent cooperative phenomenon (long range correlations) at its phase transition which is characterised by scale free behaviour. Therefore, the Ising model is critical at its phase transition. Only at this critical state, complex systems and equilibrium systems are said to exhibit similar behaviours [27, 53], thus looking into this simple Ising model can help gain insight into the complexities of a the brain.

Briefly, the Ising model is a model comprised of sites on a lattice, where each site can only be in two possible states, either up (+1) or down (-1). In this thesis, we restrict the interaction of the sites to only its nearest neighbour (in two dimensions this will be nodes to the north, south, east and west). The effective interaction strength increases or decreases depending on temperature that effects the probability of the sites being in certain configurations. Amazingly, although only nearest neighbour interaction occurs, a specific site is able to influence other sites across the entire system at the critical temperature. The existence of this long range correlation at its phase transition, makes the model very interesting despite of its simplicity.

In fact, Fraiman et al. [38] compared the Ising model at criticality and the resting state of the brain obtained from functional magnetic resonance imaging (fMRI) data sets, and concluded that both the systems exhibit identical dynamics and statistical properties. They

proposed that resting state (i.e the states when there are no explicit inputs or outputs) could be the default mode of the brain where certain areas are automatically activated during rest time and deactivated as soon as the slightest task is engaged upon. They even went further to suggest that the global changes in the brain (mood, attention etc) could be brought about the same way temperature drives the Ising model to criticality.

1.1.3 The brain operating near criticality

At first glance it looks like the brain must be critical since it has the capacity to allow communication between different parts of the brain at a speed that seems instantaneous. A critical system is highly susceptible to local distortion which resonates throughout the system. The brain seems to be doing exactly this when it switches between connections to figure out the appropriate responds to an external stimuli. Even when it is not stimulated it is perpetually changing [6, 20]. Moreover, if the brain was subcritical then only local correlations will occur and cooperative behaviour cannot be possible as the signals will die out. On the other hand, if the brain was supercritical there will be chaotic disorder as neurons will be firing and be massively activated all the time. Thus, it can be argued that the brain has to be critical for information to propagate and be comprehensible. There are even some scale free indicators found in the brain to support this hypothesis of criticality and fractal nature of the brain [25, 34, 36].

The idea that in order to be a dynamical system, the brain would have to operate near criticality was put forward by Turin in 1950 [104]. Being near criticality is very logical in terms of efficiency. Making use of local and global interaction appropriately as needed to balance out the integration and segregation as required is efficiency at its best. Brain scientists have confirmed that interactions in the brain are predominantly local [92]. Nevertheless, it does not work alone in its locality and needs to able to attain criticality very quickly. The brain connects and reconnects all the time according to its need and amazingly although the cortex is mainly excitory network it does not explosively propagate and still transmit information [24]. Thus it has to be near criticality. In fact Tagliazucchi et al. [97] claims to have evidence proving that in resting state, the brain spends most of its time near critical point where the dynamics are close to phase transition with long range correlations.

The main reason for this is that near the critical point, the system is anticipating criticality therefore a lot of different meta-stable states exist [38]. This makes the brain elastic and enhances the plasticity of the brain, to take the form of whatever state required.

One thing that we are very confident about is the fact that the brain handles loads of information. Thus some parts of the brain would be dependent on others at certain points in time. It is this flow of information and dependencies that we hope to understand a bit more of in order to contribute towards our understanding of the inner workings and complexities of the brain.

1.2 Peeking into the brain

The advances of technology have given us glimpses of what we hope to be the inner workings of the brain. Unfortunately making sense of these glimpses, is not an easy task to say the least. Adding to the difficulties is the fact that, on one hand it seems that different parts of the brain are activated at different times but on the other hand it appears almost like information is reaching all neurons at all times. Consequently, tracing the information flow in the brain will require high temporal resolution in order to capture split second changes. Electroencephalograph (EEG) may be able to do just that. On July 6, 1924 Hans Berger first used EEG to measure human brain waves [43, 84]. EEG is the recording of electrical activity along the scalp. The EEG can be defined roughly as the mean electrical activity of certain sites on the scalp [84]. The electrical activity is detected as the difference in potential between two electrodes in a grounded system. With several electrodes on the scalp, an estimate map of the brain's electrical charges can be constructed. This noninvasive technique is still the most widespread method used in laboratories [20].

Neuroimaging techniques are techniques used to visualize brain activity. The most popular of these techniques are functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and single-photon emission computed tomography (SPECT) which assess metabolic correlates of neurons in blood flow, blood volume and oxygenation respectively [20]. The most prominent among those is the fMRI due to the prevalance of MRI scanners. This method which makes use of a magnet weighing several tonnes, depends on the magnetic moments of metal ions in our blood. fMRI makes use of the fact

that the magnetic property of hemoglobin changes with presence and absence of oxygen. The information gathered by an fMRI scanner is usually processed to construct images of brain activities [20, 43]. There are many interesting studies that have made use of data sets produced by fMRI [38, 36, 97].

1.2.1 EEG versus fMRI

The strength of fMRI lies in the spatial resolution. The easy-on-the-eye resulting images obtained from the MRI scans make it possible to detect cellular activity in structures that do not contribute to scalp EEG. A main technical drawback of the fMRI method is its slow temporal resolution [20]. EEG on the other hand has a very high temporal resolution but lower spatial resolution. EEG has come a long way since the time of Hans Berger but the general idea of recording electrical activity on the scalp remains. Fortunately for us the temporal resolution now is very high to the point where we could even possibly detect reactions at a neuronal level.

EEG and fMRI are both non-invasive procedures, however there are plenty of more-than-invasive experiments performed on the brains of animals including rats [12], cats [22] and monkeys [67]. In all of these experiments the neurons reacted in order of milliseconds. Therefore we believe that the human brain reacts similarly, that cooperative neuronal interactions are manifested in the order of milliseconds which can be detected by the EEG but not the fMRI. The temporal resolution is the most important reason we are interested in EEG data sets, especially when the brain appears to reorganize itself almost instantaneously [20], in what appears to be the most effective information dissemination process in the universe. Additionally from a theoretical point of view, the high temporal resolution renders more data per second which leads to a much better probability estimate when one takes the average over time relative to the estimates that can be obtained from fMRI data sets.

Moreover, EEG technology has its advantages over the neuroimaging technique especially in terms of portability, availability and at least 85 year history of investigation. MRI machines are typically huge and noisy enclosures which can sometimes intimidate whereas EEG involves wearing a string of electrodes on ones head. It is simply psychologically and

physically more convenient. Not to mention that it is also much cheaper than the fMRI. The simplicity and widespread usage of the EEG gives it the edge. It is much easier for a pianist for example to play the piano with some electrodes on his head then in an MRI scanner. Furthermore, the existence of smaller and more efficient EEG recording machines makes data acquisition less time consuming. In our case, EEG refers to the recording of the brain's spontaneous electrical activity over a short period of time, as recorded from eight electrodes placed on the scalp.

1.2.2 Details of the EEG data set

Thanks to the team of EEG experts led by Björn Crüts at Biometrisch Centrum (BMC) [1], that make their own EEG recording devices, we have EEG data sets from possibly the most up-to-date machine. Björn's team have kindly shared some of their data sets and also gave very valuable advice with regards of the outcome of the analysis. The team is continuously working with clinicians all over the Netherlands (and currently expanding to some parts of Germany) in using EEG in psychological treatments and providing EEG machinery as well as training to the psychologist. The team strive to use the best quality materials for their machines and they are constantly trying to improve the efficiency and design of their EEG machines. The data sets provided to us were recorded at 250Hz (4 millisecond intervals) with a resolution smaller than 0.1 millivolts. These data sets motivated us to think about how the brain operates.

1.3 Measures on EEG data sets

In order to make sense of all the acquired EEG data obtained using these technologically advance machines, a certain type of measure must be used to quantify these data. There are many to choose from. The ever popular correlation deserves our attention first. But before that, some definitions that will be adopted throughout the thesis.

1.3.1 Probability and Expectation

We shall define P as the probability. The most common way we shall use P is in terms of expressing variables and its different possible values, that we shall call states. For example, define two variables X and Y that can be in states $x \in \mathcal{X}$ and $x \in \mathcal{Y}$ respectively. $\mathcal{X}$ is the state space (set of all possible states) of X and $\mathcal{Y}$ is the state space of Y. $P(X = x) = p_X(x)$ is the probability of X being in state x. $P(X = x, Y = y) = p_{XY}(x, y)$ is the joint probability of X being in state x and Y being in state y simultaneously. With regards to the conditional probability, we shall use $P(X = x|Y = y) = p_{X|Y}(x|y)$ to denote the probability of X being in state x given that Y is in state y. These terms can be generalized for relationship between many variables.

Thus if $X \sim p_X(x)$ [X has distribution $p_X(x)$] as we have defined above, then the expected value of random variable $f(X)$ is written [29] as

$$E_{p_X(x)}[f(X)] = \sum_{x \in \mathcal{X}} p_X(x) f(x) = E[f(X)]. \tag{1.1}$$

The last term on the right hand side (RHS) will be used when the probability distribution is understood from the context. Moreover, sometimes we shall take the liberty of simply using $P(X) = P(X = x)$, $P(X|Y) = P(X = x|Y = y)$ and so on, especially when using P in tandem with expectation E and when the context is clearly understood. However this is not to be confused with the general usage of P as the probability of an event. For example if $A = \{X \text{does not change}\}$ then $P(A) = P(X \text{does not change})$ is the probability that X does not change which is equal to the probability of event A.

1.3.2 Covariance and Correlation

The most common measure on neuro data sets (or any kind of data in statistical mechanics for that matter) is correlation. Correlation in a general sense of the word is commonly used to refer to the mutual relationship or connection between two or more things. Even when discussing the relationship between variables, correlation is often referred to in the context of whether or not there exist co-relation between them. In statistical mechanics, correlation often refers to a measure of mutual order existing between variables [109].

Unfortunately (and this is where it sometimes get confusing) this is usually done using the covariance measure. Recall the two variables X and Y previously defined. Given that the joint probability is $p_{XY}(x, y)$, the covariance of X and Y is defined as

$$\begin{aligned}\Gamma(X, Y) &= E[XY] - E[X]E[Y] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{XY}(x, y) - \sum_{x\in\mathcal{X}} xp_X(x)\sum_{y\in\mathcal{Y}} yp_Y(y).\end{aligned} \tag{1.2}$$

In some literature this quantity is known as the correlation, we cautiously use the term covariance here to distinguish it from the statistical correlation [36, 109],

$$\rho(X, Y) = \frac{\Gamma(X, Y)}{\sigma(X)\sigma(Y)} = \frac{E[XY] - E[X]E[Y]}{\sigma(X)\sigma(Y)} \tag{1.3}$$

where $\sigma(X)$ is the standard deviation of X (note that ρ is undefined when X or Y is constant since $\sigma(X) = 0$). The variables X and Y are said to be 'uncorrelated' or linearly independent when $\Gamma(X, Y) = 0$.

Let X^τ be the variable X that is shifted by τ time steps. For example, if X is in state 1 at time step 1, state 2 at time step 2 and so on. Denote X_n for value of X at time step n. X^τ is the variable X shifted τ time steps so that the values of X^τ are τ time step ahead of X i.e. $X^\tau_n = X_{n+\tau}$. The cross correlation is used to measure correlation across time step such that

$$\rho(Y, X^\tau) = \frac{E[YX^\tau] - E[Y]E[X^\tau]}{\sigma(Y)\sigma(X^\tau)}. \tag{1.4}$$

Similarly, cross covariance is defined as $\Gamma(Y, X^\tau)$. Autocorrelation is simply the correlation of a variable with a time shifted version of itself (hence the word auto) such that $\rho(X, X^\tau)$. Correlation is widely used in measuring quantities in many fields including neuroscience. It is extensively used on EEG data sets [5].

1.3.3 Other measures

Measures can be used in time domain or in frequency domain. The previous measures were mentioned in the context of time domain. There are many other measures that requires the

EEG data set to be in the frequency domain since there are strong reasons to believe that frequencies are important on EEG data sets. The most basic one would be coherence [44, 18], which is simply correlation as a function of frequencies allowing spatial correlation to be studied between different frequencies. For a review on how coherence have been used on EEG see [84]. More recently, popular measures for the frequency domain in neuroscience include Partial Directed Coherence (PDC) and Direct Transfer Function (DTF) by Takahashi et al. [99].

However, in this thesis we only discuss measures in the time domain. Most of our findings are similarly applicable to the frequency domain by simply replacing the probability density function with the spectral density function. We will be focusing on entropy based measures originating from information theory. The use of measures from information theory is aptly appropriate considering that it has been said to be a natural tool set to use on the brain [98] due to the vast amount of information processed.

1.4 How we humans view the brain

The way we humans picture the brain is constantly evolving. The ancient Egyptians apparently did not consider the brain to be important, thus taking it out and throwing it away from their dead ones during mummification. The ancient Greeks are the ones usually credited to be the first to think that the brain was important. The idea that the mind and matter is connected, is known to be propounded by Pythagoras who thought that the mind is somehow attuned to the laws of mathematics. Whether or not this happened in the brain was probably not important to him. However, Hippocrates who came decades later argued that the brain is the most important organ for sensation, thought, emotion and cognition [20]. Furthermore, he divided the brain into four humours responsible for different temperaments. Recent developments in technology and complexity science has resulted in paradigm shifts in terms of how we look at the brain, most of the neuroscientists are now of the view that the whole brain is also integrated in addition to having different functionalities [92].

1.4.1 Simplistic view of the brain

The idea that there exist dedicated parts of the brain for certain functionalities persisted. In the current literature it is known as functional segregation which refers to specialized neurons responding to specific input features. There are even claims that every cognitive act has its specific assembly that causes its emergence [24]. Its counterpart, the functional integration on the other hand, refers to the establishment of a statistical relationship between different and distant cell populations which may ironically degrade specialization [92]. From a very machine-like point of view the brain performs two very important and almost contradictory actions. The first is the need to extract the information from the input (stimulus) that it receives. This is where segregation is needed. At the same time (or within a few milliseconds) it has to make sense and then react to the stimulus accordingly. And this is where integration comes into play. Looking at the brain as a complex system that is constantly near criticality brings these two functions together.

Nonlinearity is fundamental in any complex system. By definition, the emergence and criticality expected in a complex system is a nonlinear phenomenon. The fact that we believe the brain is operating near criticality thus being crucially nonlinear encouraged us to look for measures that can capture nonlinearities. Nonlinear relationships are prevalent in the brain, small inputs can stimulate large outputs or sometimes none at all [20]. The functional integration (connectivity) of the brain which is defined as the connections between distant groups of cells are commonly established using temporal correlations (cross correlation by our definition) or temporal covariance [92]. We are hoping to look at things in a more nonlinear manner. Therefore, in our simplistic approach to understanding the brain, nonlinearity will be one of the main issues.

The idea of functional segregation can sometimes be explained in terms of certain specific parts of the brain controlling certain specific functions. However, these controlling areas are not the only ones that are active during the operation of the task and sometimes the communicating areas are not even spatially connected [98]. Therefore, there must be some form of communication between the areas especially when reacting to external inputs. The communication links exist mostly as a series of action potentials forming a connection. The length of this connection in a single human brain is said to be between $100,000$

to $10,000,000$ kilometres [92]. This emphasizes the importance of communication in the brain in order for it to function as a directed network. The key word here is directed, having a starting point and an end point. The very definition of the functional integration was to establish a sort of temporal relationship between groups of cells and therefore establishing 'causal' (directed) link between these groups. Therefore, the second element that we wish to incorporate in our simplistic model of the brain is a type of 'causality'.

Chapter Summary

We have first explored how the brain could be the very definition of a complex system and how technology can help us understand it better. Subsequently, we explained our interest in EEG data sets and went through some basic introduction on how people go about measuring it. The very nature of the brain where there is bound to be 'causal' connections in a very nonlinear environment of a complex system points out the need for a measure that can incorporate both nonlinearity and 'causality'. We start by examining how to capture the nonlinearity and this is where Mutual Information comes into the picture.

Chapter 2

Mutual Information as a nonlinear measure

To capture the nonlinearities of the brain, we need a nonlinear measure. The Mutual Information which is based on information theoretic entropy introduced by Shannon in his seminal paper [91] seems to fit this description. Firstly, we define entropy and its uniqueness in defining uncertainties. Consequently we define Mutual Information and Kullback Leibner measure as function of probabilities. We discuss some of the common properties of these measures. In terms of independence, we show how Mutual Information is tailor made to capture the very definition of independence in contrast to covariance which is essentially defined to capture linear independence. Lastly, we will take a brief look at some applications of the Mutual Information including possible 'causality' detection.

2.1 Entropy

In statistical mechanics, entropy arises as a measure of uncertainties and disorganization in a physical system. Information theory deals with entropy in a slightly different manner as it focuses on the quantification of information. Mutual Information is an entropy based measure widely used in information theory. In order to understand Mutual Information, we need to understand entropy. Entropy was first coined in 1865 by Rudolf Julius Emmanuel Clausius [59] in reference to thermodynamics. However in 1948, Shannon de-

fined a slightly different kind of entropy in terms of measurable dynamics and information theory. It was reported that he contemplated calling it information or uncertainty but was convinced by John Von Neumann to call it entropy for two reasons. The first was that entropy is the uncertainty function of statistical mechanics and secondly according to Neumann, to call it entropy will be advantageous in debates since nobody knows what entropy really is [59].

On a more general level, the field of complex networks is still looking for a good quantifier of complexity and there seems to be a great need for a new theory of information of complex networks [3]. One way to tackle this issue is to import the key concepts of information theory that has quantification of information as the main focus where entropy based measures such as Mutual Information plays a key role. Being able to evaluate the information transfer of a complex system is one of the main outstanding problems in the statistical mechanics of the complex systems.

2.1.1 The definition of entropy

Shannon [91] defined entropy as the function $H(p_1, p_2, ..., p_n)$ for probabilities $p_1, p_2, ..., p_n$,

$$H(p_1, p_2, ..., p_n) = -k \sum_{i=1}^{n} p_i \, log \, p_i \tag{2.1}$$

where $k > 0$ is a constant and the log is to the base 2. The aim was to measure the uncertainty of the outcome of a certain variable given the probabilities. H was chosen to represent uncertainty due to it being the only the solution to certain properties that are outlined to 'measure' uncertainty [91] . Khinchin [58] came up with a more mathematically rigorous proof for uniqueness of H for the chosen properties. Before outlining these properties in subsection (2.1.2), we explain more about H.

When $p_1, p_2, ..., p_n$ are the probabilities of discrete random variable X we write

$$H(X) = H(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \, log \, p_i \tag{2.2}$$

as the entropy of X [29, 60, 103]. Without lost of generality we have set $k = 1$. Note that

X is not the argument of the function H but it is the random variable the entropy of which we are trying to measure. For example, say we have a random variable X with probabilities $p_1 = P(X = 1) = \frac{1}{3}$, $p_2 = P(X = 2) = \frac{1}{2}$ and $p_3 = P(X = 3) = \frac{1}{6}$, then the entropy would be

$$H(X) = -\frac{1}{3}\, log\, \frac{1}{3} - \frac{1}{2}\, log\, \frac{1}{2} - \frac{1}{6}\, log\, \frac{1}{6}.$$

Entropy does not actually depend on the values taken by random variable X but only its probabilities. This entropy definition is also prominent in measure theory and ergodic theory [52, 59].

More formally, define two random variables X and Y that have probability $p_X(x), x \in \mathcal{X}$ and $p_Y(y), y \in \mathcal{Y}$, respectively. $\mathcal{X}$ is the state space (set of all states) of X and $\mathcal{Y}$ is the state space of Y. Their respective entropies [29] are

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) log\, p_X(x) \tag{2.3a}$$

and

$$H(Y) = -\sum_{y \in \mathcal{Y}} p_Y(y) log\, p_Y(y). \tag{2.3b}$$

We will use the convention $0\ log\ 0 = 0$ which is easily justified by continuity since $x\ log(x) \to 0$ as $x \to 0$ and log to the base e. In terms of E, we have

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) log\, p_X(x) = E_{p_X(x)}\left[-log\, p_X(x)\right] = E\left[-log\ P(X)\right]. \tag{2.4}$$

Therefore, the entropy of X can also be interpreted as the expected value of $-log\, p_X(x)$ when $X \sim p_X(x)$ [29]. Recall that $p_X(x) = P(X = x)$ and E is the expectation as previously defined in subsection (1.3.1).

2.1.2 The uniqueness of entropy

When more than one variable is involved, joint and conditional entropy can be defined [29, 52]. The joint entropy of X and Y is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y) log\, p_{XY}(x,y) = E\left[-log\, P(X,Y)\right]. \tag{2.5}$$

Define $H(Y|X=x) = -\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x) log\, p_{Y|X}(y|x)$, so that the conditional entropy can be written as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p_X(x) H(Y|X=x) = -\sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{Y}} p_{Y|X}(y|x)\, log\, p_{Y|X}(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\, log\, p_{Y|X}(y|x) = E_{p_{XY}(x,y)}[-log\, P(Y|X)] \end{aligned} \tag{2.6}$$

where we have substituted $p_{Y|X}(y|x)p_X(x) = p_{XY}(x,y)$.

The Uniqueness of Entropy was later on proven in mathematically rigorous way by Khinchin [58] when H has these properties:

1. Given n possibilities, H is maximum when $p_X(x) = \frac{1}{n}$ for $\forall x \in \mathcal{X}$.
2. The chain rule: $H(X,Y) = H(X) + H(Y|X)$.
3. Adding an impossible event i.e a zero probability event does not change the value of H: $H(p_1, p_2, ..., p_n, 0) = H(p_1, p_2, ..., p_n)$.

As a side note, we would like to point out a trivial point that will make working with expectations much simpler. For any function of X, $f(X)$ with distribution function $p_X(x)$, we have that

$$E_{p_X(x)}[f(X)] = \sum_{x\in\mathcal{X}} p_X(x) f(x) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y) f(x) = E_{p_{XY}(x,y)}[f(X)], \tag{2.7}$$

where we made use of the joint probability property $\sum_{y\in\mathcal{Y}} p_{XY}(x,y) = p_X(x)$. Consequently, the chain rule on H (property 2) can be verified by substituting the Bayes theorem

$P(X,Y) = P(X)P(Y|X)$ into equation (2.5) as follows

$$\begin{aligned} H(X) + H(Y|X) &= E_{p_X(x)}[-log\ P(X)] + E_{p_{XY}(x,y)}[-log\ P(Y|X)] \\ &= E_{p_{XY}(x,y)}[-logP(X)P(Y|X)] = E_{p_{XY}(x,y)}\left[-log\ P(X,Y)\right] \\ &= H(X,Y). \end{aligned} \tag{2.8}$$

This rule links joint and conditional entropy. By symmetry we also have $H(X,Y) = H(Y) + H(X|Y)$. It is possible to include more variables in the definition of joint and conditional entropy where the definitions can also be linked using the chain rule. If we define a new random variable Z in a similar manner, using the equality from Bayes theorem $p_Z(z)p_{XY|Z}(x,y|z) = p_{XYZ}(x,y,z)$ and the definition of the conditional entropy, we get

$$\begin{aligned} H(X,Y|Z) &= \sum_{z\in\mathcal{Z}} p_Z(z)H(X,Y|Z=z) \\ &= -\sum_{z\in\mathcal{Z}} p_Z(z) \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY|Z}(x,y|z)\, log\, p_{XY|Z}(x,y|z) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{XYZ}(x,y,z)\, log\, p_{XY|Z}(x,y|z) \\ &= E_{p_{XYZ}(x,y,z)}\left[-log\ P(X,Y|Z)\right]. \end{aligned} \tag{2.9}$$

And by substituting $p_{YZ}(y,z)p_{X|YZ}(x|y,z) = p_{XYZ}(x,y,z)$, we obtain

$$\begin{aligned} H(X|Y,Z) &= \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{YZ}(y,z)H(X|Y=y,Z=z) \\ &= -\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{YZ}(y,z) \sum_{x\in\mathcal{X}} p_{X|YZ}(x|y,z) log\, p_{X|YZ}(x|y,z) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{XYZ}(x,y,z) log\, p_{X|YZ}(x|y,z) \\ &= E_{p_{XYZ}(x,y,z)}\left[-log\ P(X|Y,Z)\right]. \end{aligned} \tag{2.10}$$

It is clear that by simple manipulation of the probabilities using Bayes theorem we can get relations between conditional entropies. For example, using Bayes theorem to manipulate

the probabilities we get

$$p_{XY|Z}(x,y|z) = \frac{p_{XYZ}(x,y,z)}{p_Z(z)} = \frac{p_{Y|XZ}(y|x,z)p_{XZ}(x,z)}{p_Z(z)}$$
$$= p_{Y|XZ}(y|x,z)p_{X|Z}(x|z),$$

so that the entropy of X and Y conditioned on Z is equal to entropy of Y conditioned on X and Z plus the entropy of X conditioned on Z, i.e.

$$H(X,Y|Z) = H(Y|X,Z) + H(X|Z). \tag{2.11}$$

We are now prepared to define Mutual Information.

2.2 Mutual Information and Relative Entropy

For the same random variables X and Y, the Mutual Information [29, 60, 61, 64, 73, 103] is defined as

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= E_{p_X(x)}[-log\ P(X)] - E_{p_{XY}(x,y)}[-log\ P(X|Y)] \\ &= -E[log\ P(X)] + E_{p_{XY}(x,y)}\left[log\frac{P(X,Y)}{P(Y)}\right] \\ &= E_{p_{XY}(x,y)}\left[log\frac{P(X,Y)}{P(X)P(Y)}\right] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \end{aligned} \tag{2.12}$$

I can be interpreted as the amount of information Y provides about X since it measures the difference between the uncertainty of X and the the uncertainty of X given Y. The relationship can also be viewed in a set theoretic setting as in Figure (2.1). If we define $H(X)$ and $H(Y)$ as sets, then we have that $I(X,Y) = H(X) \cap H(Y)$ and that $H(X,Y) = H(X) \cup H(Y)$. The conditional entropies are given by $H(X|Y) = H(X) \setminus I(X,Y)$ and $H(Y|X) = H(Y) \setminus I(X,Y)$. From the definition and the Venn diagram we can see that I is

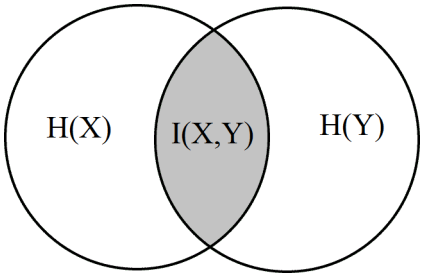

Figure 2.1: Venn diagram depicting relationship between I and H

symmetric so that the information that Y gives about X is the same amount of information X gives about Y. The symmetry of Mutual Information can be shown using the chain rule in equation (2.8).

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) = H(X) - [H(X,Y) - H(Y)] \\ &= H(Y) - [H(X,Y) - H(X)] = H(Y) - H(Y|X) = I(Y,X). \end{aligned} \tag{2.13}$$

Note that $I(X,X) = H(X) - H(X|X) = H(X)$ since $H(X|X)$ is obviously zero. Therefore H is a special case of I.

2.2.1 Relative Entropy and conditional Mutual Information

Another interesting way of interpreting Mutual Information is through relative entropy. Relative entropy or Kullback Leibner (KL) distance [29, 52, 63] between two distribution functions $f(x)$ and $q(x)$ is defined as

$$D(f||q) = \sum_{x \in \mathcal{X}} f(x)\, log\, \frac{f(x)}{q(x)} = E_f \left[log\, \frac{f(X)}{q(X)} \right],$$

where we take $0\, log\, \frac{0}{q} = 0$ and $0\, log\, \frac{f}{0} = \infty$. D can be seen as a measure of distance between distributions. However it should be pointed out that it is not a true metric distance because it is not symmetric and does not satisfy the triangle inequality. Despite that, it is often helpful to think of D as the distance between f and q, because it seeks to quantify the difference between these two distributions and $D = 0$ when $f(x) = q(x)$. If we let

$f(x) = p_{XY}(x, y)$ and $q(x) = p_X(x)p_Y(y)$ then we get that

$$D(p_{XY}(x,y)||p_X(x)p_Y(y)) = E_{p_{XY}(x,y)}\left[log\,\frac{P(X,Y)}{P(X)P(Y)}\right] = I(X,Y). \tag{2.14}$$

It appears that I is just the special case of D where f is the joint distribution and q is the product distribution. Using this logic, we can interpret I as measuring how 'far' the joint distribution is from the product distribution.

Analogous to entropy, the definition of Mutual Information can be extended to include a third random variable Z in the form of conditional Mutual Information [29, 52] which can be written as

$$\begin{aligned} I(X,Y|Z) &= H(X|Z) - H(X|Y,Z) = E[-log\,p(X|Z)] - E[-log\,p(X|Y,Z)] \\ &= E\left[log\,\frac{p(X|Y,Z)}{p(X|Z)}\right] = E\left[log\,\frac{p(X,Y,Z)p(Z)}{p(Y,Z)p(X,Z)}\right] \\ &= E_{p_{XYZ}(x,y,z)}\left[log\,\frac{P(X,Y|Z)}{P(Y|Z)P(X|Z)}\right], \end{aligned} \tag{2.15}$$

where one can clearly see that the only difference with the Mutual Information definition in equation (2.12) is that the probabilities are conditioned on Z. Conditional Mutual Information can be interpreted as the amount of information X provides about Y (or vice versa) given that Z is known.

2.2.2 Properties of Entropy, Relative Entropy and Mutual Information

Entropy H is a specific case of Mutual Information I. Mutual Information I is a specific case of relative entropy D. So all properties of D extends to I and all properties of I extends to H. This does not apply in the other direction. One prevailing property for all three quantities is non-negativity. This can be proven using Jensen's Inequality [29, 108] which states that if g is a convex function then $E[g(X)] \geq g(E[X])$. And if g is a concave function then $E[g(X)] \leq g(E[X])$. Let $f(x)$ and $q(x)$ be distribution functions, we claim that $D(f||q) \geq 0$ and $D(f||q) = 0$ if and only if $f(x) = q(x)$ for any x. For $f(x), q(x) > 0$

we have that

$$\begin{aligned}-D(p||q) = -E_{f(x)}\left[log\frac{f(X)}{q(X)}\right] &= E_{f(x)}\left[log\frac{q(X)}{f(X)}\right] \\ &\leq log\, E_{f(x)}\left[\frac{q(X)}{f(X)}\right] = log\sum_x f(x)\frac{q(x)}{f(x)} = log\sum_x q(x) = log 1 = 0\end{aligned}$$

where the inequality follows from Jensen's inequality since the function $log(x)$ is a concave function of x. Thus $I(X,Y) = D(p_{XY}(x,y)||p_X(x)p_Y(y)) \geq 0$ and $I(X,Y) = 0$ if and only if $p_{XY}(x,y) = p_X(x)p_Y(y)$. Moreover $H(X) = I(X,X) \geq 0$ and $H = 0$ only when X is constant [29]. However it must be pointed out that

$$\begin{aligned}I(X,Y|Z) &= E_{p_{XYZ}(x,y,z)}\left[log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] \\ &\neq E_{p_{XY|Z}(x,y|z)}\left[log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] \\ &= D(p_{XY|Z}(x,y|z)||p_{X|Z}(x|z)p_{Y|Z}(y|z)),\end{aligned}$$

and that $D \geq 0$ does not necessarily implies that the conditional Mutual Information $I(X,Y|Z) \geq 0$. However $I(X,Y|Z) \geq 0$ can also be proven using Jensen's inequality.

$$\begin{aligned}I(X,Y|Z) = E_{p_{XYZ}(x,y,z)}\left[log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] &= E_{p_{XYZ}(x,y,z)}\left[-log\frac{P(X,Z)P(Y,Z)}{P(X,Y,Z)P(Z)}\right] \\ &\geq -log\left[E_{p_{XYZ}(x,y,z)}\frac{P(X,Z)P(Y,Z)}{P(X,Y,Z)P(Z)}\right] \\ &= -log\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{XYZ}(x,y,z)\frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_{XYZ}(x,y,z)p_Z(z)} \\ &= -log\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\frac{p_{YZ}(y,z)}{p_Z(z)}\sum_{x\in\mathcal{X}} p_{XZ}(x,z) = -log\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\frac{p_{YZ}(y,z)}{p_Z(z)}p_Z(z) \\ &= -log\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p_{YZ}(y,z) = -log\sum_{z\in\mathcal{Z}} p_Z(z) = -log\ 1 = 0.\end{aligned}$$

One property of I not possessed by D is symmetry. Generally, D is not symmetric since $D(p||q) = E_p log\frac{p(X)}{q(X)} \neq E_q log\frac{q(X)}{p(X)} = D(q||p)$. I however, is symmetric because of

equation (2.13). Joint entropy $H(X,Y)$ is symmetric since we get that

$$\begin{aligned} H(X,Y) &= E_{p_{XY}(x,y)}[-log\ P(X,Y)] = H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) = H(Y,X). \end{aligned}$$

On the other hand, conditioning reduces entropy hence distorting the symmetry i.e

$$\begin{aligned} H(X|Y) &= -E_{p_{XY}(x,y)}[log\ P(X|Y)] = H(X,Y) - H(Y) \leq H(X), \\ H(Y|X) &= -E_{p_{XY}(x,y)}[log\ P(Y|X)] = H(X,Y) - H(X) \leq H(Y), \end{aligned}$$

so that $H(X|Y) \neq H(Y|X)$ in general. It is the same case for I, generally $I(X,Y|Z) \neq I(X,Z|X) \neq I(Y,Z|X)$ since the probabilities could be different although we obviously still have $I(X,Y|Z) = I(Y,X|Z)$.

2.3 Mutual Information versus covariance

One feature of Mutual Information is that it enables quantification of relationship between symbolic sequences [64]. This is due to the fact that Mutual Information only depends on probabilities rather than the values of the variable itself, consequently the variable does not have to be a number. It is pointed out that, in this way Mutual Information is somewhat more flexible than covariance (and correlation) function since it does require the variable to be a number. In a similar sense we can get relationship between blocks or group of nodes on the brain since all we need are the probabilities and not the values of the nodes itself. This could be very useful in identifying any scale free features of the brain to support the claim in [36]. However the most important difference is in terms of linearity and independence.

2.3.1 Comparing independence

The variables X and Y are said to be linearly independent of each other if

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{XY}(x,y) = \sum_{x\in\mathcal{X}} xp_X(x)\sum_{y\in\mathcal{Y}} yp_Y(y). \tag{2.16}$$

It can easily be seen that the covariance and correlation is defined to capture linear independence. Recall the definition of covariance from equation (1.2) so that

$$\begin{aligned}\Gamma(X,Y) &= E[XY] - E[X]E[Y] = 0 \\ &\Rightarrow\ E[XY] = E[X]E[Y] \\ &\Rightarrow \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{XY}(x,y) = \sum_{x\in\mathcal{X}} xp_X(x)\sum_{y\in\mathcal{Y}} yp_Y(y)\end{aligned} \tag{2.17}$$

where the last line is the very definition of linear independence. Therefore $E[XY] = E[X]E[Y]$ implies that X and Y are linearly independent.

It is well known that two variables X and Y are independent if and only if their joint distribution equals their product distribution [108] such that

$$p_{XY}(x,y) = p_X(x)p_Y(y)\ \forall\, x\in\mathcal{X}, y\in\mathcal{Y}. \tag{2.18}$$

It is a well known fact in statistics that uncorrelated-ness (linear independence) does not imply (general) independence but independence implies uncorrelated-ness [46]. Anything that is independent would of course be linearly independent by default. Recall from equation (2.14) that when $p_{XY}(x,y) = p_X(x)p_Y(y)$ i.e when the X and Y are independent, we have that $I = 0$. We have seen that I can be interpreted as measuring how 'far' the joint distribution is from the product distribution (which is the joint distribution when X and Y are independent). In other words Mutual Information seeks to measure how dependent these two variables are on each other. One can see this from the definition of Mutual Information in equation (2.12) which renders

$$\begin{aligned}I(X,Y) = E\left[log\,\frac{P(X,Y)}{P(Y)P(X)}\right] &= E[\,log\,P(X,Y) - log\,P(X)P(Y)] = 0 \\ \Rightarrow\ E[\,log\,P(X,Y)] &= E[\,log\,P(X)P(Y)].\end{aligned} \tag{2.19}$$

Consequently it is logical to expect that, Mutual Information has the potential to provide us with insights that have not been obtained using covariance before. The direct approximation approach should give us a clear indication of independence given that the probabilities

are accurately estimated. Moreover, even if X and Y are not independent of each other, one could have

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \; \forall \, x, y, z$$

so that variables X and Y are said to be conditionally independent on variable Z. Therefore, if X and Y are conditionally independent on Z, we get that $I(X, Y|Z) = 0$.

2.3.2 An illustrative example

The discussion on independence indicates that $\Gamma = 0 \not\Rightarrow I = 0$ but $I = 0 \Rightarrow \Gamma = 0$. This is demonstrated in [64] which starts off with a binary sequence where the numerical values of X and Y can only be either 0 or 1. In this case, we can find a relationship between these two functions and linear dependence would lead to general dependence. Therefore we have that $\Gamma = 0 \Leftrightarrow \; I = 0$ for binary sequences.

Let $p(\alpha)$ and $p(\alpha, \beta)$ be the probability and joint probability for $\alpha, \beta \in \{0, 1\}$. The covariance in equation (1.2) becomes

$$\Gamma(X, Y) = E[XY] - E[X]E[Y] = p(1, 1) - p(1)^2, \tag{2.20}$$

so that we can write the probabilities in terms of $\Gamma = \Gamma(X, Y)$. Using the property of joint probability $\sum_\beta p(\alpha, \beta) = p(\alpha)$ and imposing $p(\alpha, \beta) = p(\beta, \alpha)$, we get that $p(1) - p(1, 1) = p(0) - p(0, 0)$. Moreover, taking into account $p(1) + p(0) = 1$ (the normalizing condition for probabilities) we obtain

$$p(1, 1) = \Gamma + p(1)^2, \tag{2.21a}$$

$$p(0, 0) = \Gamma + p(0)^2, \tag{2.21b}$$

$$p(0, 1) = p(1, 0) = -\Gamma + p(0)p(1). \tag{2.21c}$$

The probabilities can be used to obtain the Mutual Information formula using equation

(2.12) so that

$$\begin{aligned} I(X,Y) =& \sum_{\alpha}\sum_{\beta} p(\alpha,\beta) log \frac{p(\alpha,\beta)}{p(\alpha)p(\beta)} \\ =& \Gamma \ log \frac{\left(1+\frac{\Gamma}{p(1)^2}\right)\left(1+\frac{\Gamma}{p(0)^2}\right)}{\left(\frac{\Gamma}{p(0)p(1)}\right)^2} + p(1)^2 log\left(1+\frac{\Gamma}{p(1)^2}\right) \\ &+ p(0)^2 log\left(1+\frac{\Gamma}{p(0)^2}\right) - 2p(0)p(1) log\left(1+\frac{\Gamma}{p(0)p(1)}\right). \end{aligned} \tag{2.22}$$

This equation ties in with the fact that the lower bounds on Mutual Information I for any kind of sequence has been proven to be dependent on the covariance Γ and the marginal probabilities [37]. An approximation when the terms $\frac{\Gamma}{p(\alpha)p(\beta)}$ are small gives

$$I \approx \frac{\Gamma^2}{2}\left(\frac{1}{p(1)^2}+\frac{1}{p(0)^2}+\frac{2}{p(0)p(1)}\right) = \frac{1}{2}\left(\frac{\Gamma}{p(0)p(1)}\right)^2. \tag{2.23}$$

This illustrates that I decays to zero at a faster rate than the corresponding Γ.

By setting $\alpha, \beta \in \{0,1,2\}$, we get ternary sequences. To obtain a relationship between I and Γ using similar methods and constraints does not seem possible for ternary sequences. However, setting $\Gamma = 0$ in equation (1.2) with current values of X and Y renders

$$\begin{aligned} \Gamma &= E[XY] - E[X]E[Y] \\ &= p(1,1) + 2p(1,2) + 2p(2,1) + 4p(2,2) - (p(1) + 2p(2))^2 = 0 \end{aligned} \tag{2.24}$$

and using this as an additional constraint, gives us probabilities to put in I to get values of I corresponding to $\Gamma = 0$. In [64], some non-negative values for the probabilities were randomly chosen and this made clear that there are values for which $\Gamma = 0$ but $I \neq 0$. Therefore $\Gamma = 0 \not\Rightarrow I = 0$ for ternary sequences in general. This demonstrates that the Mutual Information function is capable of capturing the nonlinear dependencies that the covariance might have missed.

2.4 Various applications of Mutual Information

Just like correlation, Mutual Information has also been defined and redefined to serve different purposes. The most natural extension is to insert an element of time and use the Mutual Information with its time shifted counterpart ala cross correlation in equation (1.4) such that $I(X, Y^\tau) = H(X) - H(X|Y^\tau)$ where Y^τ denotes a variable Y which has been shifted by time τ. This value is known as pairwise cross Mutual Information [40] or time delayed Mutual Information [57, 89]. Mutual Information applied to its time shifted version is sometimes referred to as auto Mutual Information [2, 55, 56] such that $I(X, X^\tau) = H(X) - H(X|X^\tau)$ where X^τ is the variable X that is shifted by τ.

Some go even further by defining the persistent Mutual Information of a variable (sometimes know as Mutual Information between past and future) which is the Mutual Information of the past history of a variable and it's evolution later in the future [8]. Analytical works has been done on continuous Mutual Information for stochastic differential equations on Gaussian cases where Mutual Information has been expressed as the mean square estimation error [7, 32, 33]. Mutual Information have also been applied on the frequency domain [18] and is said to be better than coherence.

2.4.1 Quantifying transitions

Aiming to predict the future of evolving dynamical systems from the past using observed historical data, [8] uses the persistent Mutual Information on the logistic map and concluded that the measure succeeded in detecting different types of associated cascades of banded chaos in addition to period doubling. This is an example of how Mutual Information is utilized to quantify underlying transition in dynamical systems.

In [73], it is claimed that the Mutual Information is able to detect the phase transition occurring in a two dimensional Ising model. This claim has been corroborated on different systems on a few occasions. On the Viscek model of self propelled particles for example, [105] has claimed that Mutual Information is a sensitive indicator and phase transition locator. Furthermore, [105] claims that on this particular model, the Mutual Information works even better than susceptibility even when only partial observations are available. Drawing parallels between market crashes and phase transition under the assumption that

collective pricing behaviour of the financial market works the like a complex systems, [47] claims that Mutual Information indicates the transition from random to collective behavior on the data sets.

2.4.2 Clustering and hierarchy detection

There are strong indicators that there exist hierarchies in the brain [20]. The grouping property of Mutual Information could provide a natural way for application in clustering algorithms. For this purpose, one could adopt the definition in [61] such that

$$MI(X,Y,Z) = H(X) + H(Y) + H(Z) - H(X,Y,Z)$$

is defined for random variables X, Y, Z as utilized before. This representation is useful since one could make use of the grouping property of Mutual Information so that

$$MI(X,Y,Z) = I(X,Y) + I((X,Y),Z),$$

where $I((X,Y),Z) = I(X,Z) + I(Y,Z|X)$ [29, 61]. If we again define everything in set theoretic terms like in Figure (2.1) and consider $H(X)$, $H(Y)$ and $H(Z)$ as sets so that $I(X,Y) = H(X) \cap H(Y)$, $I(Y,Z) = H(Y) \cap H(Z)$ and $I(X,Z) = H(X) \cap H(Z)$, then

$$MI(X,Y,Z) = I(X,Y) \cup I(Y,Z) \cup I(X,Z) \tag{2.25a}$$

and

$$I((X,Y),Z) = I(X,Z) \cup I(Y,Z). \tag{2.25b}$$

We can clearly see from Figure (2.2) that

$$MI(X,Y,Z) = I(X,Y,Z) + I(X,Y|Z) + I(Y,Z|X) + I(X,Z|Y). \tag{2.26}$$

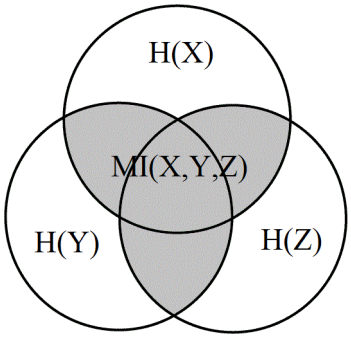

Figure 2.2: $MI(X, Y, Z)$, $H(X)$, $H(Y)$ and $H(Z)$

These equalities can be generalized into

$$MI(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \cdots, X_n)$$

$$I((X_1, X_2, \cdots, X_{n-1}), X_n) = \sum_{i=1}^{n} I(X_i, X_n | X_1, \cdots, X_{i-1})$$

where $H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1})$ is the generalization [29] of the chain rule on entropy that we have defined before. Therefore Mutual Information could also be used as the proximity measure in a clustering algorithm, where the equations can be used recursively as a clustering tool. This have been done in [61] electrocardiogram (ECG) data. Examples of Mutual Information used for clustering and classification of EEG data are given in [56, 2, 55]. Other applications of Mutual Information based clustering can be seen in [21, 4].

2.4.3 Detecting causality

The earliest criticism of ‘causality’ based test is that correlation is not equal to ‘causality’ due to its symmetry [46]. Correlation and Mutual Information gives no indication of the direction of the relationship. Coming back to our aim of capturing both nonlinearity and ‘causality’, it would be great if Mutual Information is able to quantify ‘causality’. Therefore one would logically conclude that inserting an element of time into conditional Mutual Information will be more suitable for ‘causality’ detection. It so happens that there is a value called Transfer Entropy [89] that is said to be able to do just that. This value is

simply an extension of the conditional Mutual Information where the conditional Mutual Information is utilized but now with its time shifted counterpart.

Consider the time shifted variable as in equation (1.4) and let X^1 be the variable X that is shifted by 1 (so that the values of X^1 always comes before X). Recall the definition of conditional Mutual Information from equation (2.15) and let $X = X^1$ and $Z = X$ in the definition so that

$$\begin{aligned} I(X, Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(X^1|X) - H(X^1|Y, X) = I(X^1, Y|X) = T_{Y\to X}. \end{aligned} \tag{2.27}$$

This is a simple example of Transfer Entropy $T_{Y\to X}$ as described by [89]. The idea is that, if Y causes X at time lag $\tau = 1$, then $I(X^1, Y|X)$ will be large since a lot the uncertainties of X will be caused by Y and the term $H(X^1|Y, X)$ will be much smaller than $H(X^1|X)$. However, before we delve further into how exactly the Transfer Entropy works we first need to address the question of 'causality'.

Chapter Summary

Entropy based measures are at the heart of information theory. There are overwhelming interest in entropy based measure due to its nonlinearity. Moreover, we have discussed our interest in Mutual Information due to its direct approach to quantifying 'independence' as opposed to covariance. Popular applications of Mutual Information based measures and its variants includes quantifying transitions, clustering and possibly 'causality' detection. The relationship between independence and 'causality' is eminent. However, it seems that dependency needs to be coupled with some sort of time element in order to be asymmetric and therefore 'causal'. In the next section we explore the idea of 'causality' and what can be done to quantify it.

Chapter 3

The question of ‘causality’

The question of ‘causality’ is a very tricky one to say the least. The definition of it, is not something universally accepted. First we examine the conceptualization of ‘causality’ as envisioned by Wiener and formulated by the nobel prize winning Granger. It is in this framework of ‘causality’ that we wish to build upon. We shall discuss some terms and issues related to this kind of ‘causality’ for the sake of clarity. Next, we expound on what type of ‘causality’ do we expect in the brain and how we intend to detect this using asymmetric measures. The most popular ‘causality’ measure is Granger Causality (G-causality), therefore we venture into taking a closer look at G-causality and the challenges it faces. Lastly, we go through some generalizations of G-causality (which surprisingly includes Transfer Entropy) that aims to address the challenges.

3.1 The concept of ‘causality’

In his book *I Am A Mathematician* [106], Norbert Wiener wrote that

> “ .. If we can measure degrees of causality ... We can then observe how much a change in one aspect of the universe will bring out changes in others.”

Wiener implied in his speech (and later in his book and papers) that it is possible to quantify ‘causality’ by virtue of quantifying the changes in a certain variable that incites changes in another. Wiener had the idea that the ‘causality’ of a variable in relation to another can be measured by how well the variable helps to predict the other. In other words, variable

A 'causes' variable B if the ability to predict B is improved by incorporating information about A in the prediction of B. Moreover, Wiener also said in [106] that

> ".. I was forced to consider the theory of information, and above all, that partial information which our knowledge of one part of the system gives us the rest of it."

The multi-talented Wiener is considered to be one of the pioneers of information theory. He believed that information theory could really contribute to detecting 'causality' and uncovering hidden information.

3.1.1 Different point of views on 'causality'

From a philosophical point of view there has never been a clear agreement on what could be defined as 'causality', for an interesting review of the mathematical theory of causation from a philosophical point of view refer to [50]. Some philosophers even hold the view that 'causality' is impossible to quantify [45, 52].

Statisticians often meet with 'causality' when dealing with correlation coefficient and regression [45]. Granger has written a review (mainly intended for econometricians) about the concept of 'causality' [46]. A more recent overview of causal related statistics albeit in a slightly different area is written by Judea Pearl [80] who is known as one of the pioneers of the Bayesian networks. According to him, the recent statistical ideas are moving away from traditional statistical analysis and more towards causal analysis. He differentiates between these two by saying that traditional statistical analysis focuses more on describing the data and inferring distribution parameter from samples while causal analysis requires explicit articulation of the underlying causal assumptions which is not what Bayesian statistician normally do.

In Bayesian statistics (the name derived from Bayes theorem for conditional probability), graphical models are often used. Graphical models are probabilistic models denoting conditional independence structure between random variables. In [80], Pearl proposes using Structural Causal Model (SCM) to define causal quantities, causal assumptions and all the other concepts needed in a causal discourse. SCM is an extension on the Structural

Equation Modelling on linear systems. Granger admitted that it is possible to incorporate a more Bayesian viewpoint to the idea of ‘causality’ by incorporating dynamics of prior beliefs in the model [46]. One thing all the methods mentioned previously have in common is that one will first need to fit a model to the data in order to extract the ‘causality’ and that most of the models are essentially linear or at least based on a linear model. The model-free quantifications of ‘causality’ seem to have their root in information theory.

In agreement with Pearl that causal statistics is one of the most important statistics, [52] summarizes the information-theoretic and dynamical systems approach to causality. The paper explains that the link between these two fields is due to the fact that many of the approaches to inferring causality from experimental time series came about from studying synchronization of chaotic systems where the Shannon’s entropy definition has been adopted to study dynamical systems in the ergodic theory [59]. Various information-theoretic functionals have been used to estimate, classify and and explain chaotic data [8, 52].

3.1.2 The arrow of time and prediction

Despite all those differing views on ‘causality’ even the philosophers [17, 87, 50] agree on the fact that the causal variable must come before the affected variable. As far as we know, the future cannot cause the past and the arrow of time persists. Hence, there must exist a certain time lag however small between the cause and the effect, this will be henceforth referred to as the causal lag [44]. Granger himself said that the flow of time clearly plays a central role and there is no use attempting to discuss ‘causality’ without time.

Another recurring theme is the use of prediction in ascertaining whether or not the causal variable has unique information about the affected variable which implies that we can infer ‘causality’ by comparing predictions. Consequently, we outline standard steps of inferring ‘causality’ derived from Wiener’s idea, Granger’s formulation and the basic assumption that the knowledge of the causal variable helps forecast the affected variable. It is this definition of ‘causality’ that we will adopt in this thesis. Say we want to test whether variable Y causes variable X. The first step would be to predict the current value of X using the historical values of X. The second step is to do another prediction where the

historical values of Y and X are both used to predict the current value of X. And the last step would be to compare the former to the latter. If the second prediction is judged to be better than the first one, then one can conclude that Y causes X.

3.2 Issues in 'causality'

3.2.1 Directionality and information transfer

In networks literature the references to 'causality' take many guises. The term directionality, information transfer and sometimes even independence can possibly refer to some sort of 'causality' in line with our previously defined concept. 'Causality' plays a main role when [76] discusses flow of information in Bayesian systems and when [66] expounds on way to formalize information transfer on fully known dynamical systems. Our definition of 'causality' is based on how well a variable helps the prediction of another variable. Now let us assume that Y causes X.

We would expect the relationship between X and Y to be asymmetric and that the information flows in a direction from Y to X. [68, 66, 52] highlights the importance of asymmetry in information transfer. When it comes to directionality it is paramount to point out that the main reason correlation is not equal to causation is due to the fact that causation has direction and thus essentially asymmetric [44]. When one variable causes another variable obviously the affected variable depends on the causal variable. Therefore, one could also say that our prediction based definition of 'causality' is equivalent to looking for dependencies between the variables at a certain causal lag.

Information transfer needs a source and target. The source where the information is from and the target where it is transferred to. Thus in the case of 'causality' the source will be the causal variable and the target is the affected one. One can assume that this information transfer is the unique information provided by the causal variable to the affected one. However this does not mean that the causal variable has complete control over the affected variable.

3.2.2 Deterministic variables and instantaneous casuality

If a variable has complete control of another variable then it is deterministically determined by the control variable and thus indistinguishable from it. A purely deterministic variable cannot be said to have any other causal influence other than its own past and a simple example of a deterministic case given by [44] illustrates this. Let there be variables X and Y where $X_t = bt$ and $Y_{t-1} = c(t+1)$. Then X can exactly be predicted by the equation $X_t = b + X_{t-1}$ or equally by the equation $X_t = \frac{b}{c} Y_{t-1}$. The predictions using both X_{t-1} and Y_{t-1} are exactly the same hence indistinguishable. Moreover, one can also express Y_t in terms of X_{t-1} through the formula $Y_t = \frac{c}{b}(X_{t-1} + 2b)$ which is equivalent to $Y_t = Y_{t+1} + c$. Consequently it seems that 'causality' requires the variables to be stochastic. With the uncertainty one is able to measure the 'causal' element and the directionality. The 'causal' and affected variable needs to have an independent source of variation [50].

The notion of instantaneous causality is discussed in [46]. The idea that two variables can instantaneously cause each other with no causal lag at all has been said to be impossible. Granger maintains that true instantaneous causality can never occur [46] and if anything appears to be like it, then the 'causality' is either not measured at the correct time scale (the causal lag is smaller than the measured time scale) or there is another variable jointly (or indirectly) causing it which is not observed (not incorporated in the model).

3.2.3 Indirect 'causality' and independence

Granger pointed out that apparent instantaneous causality could be caused by variables that were not incorporated in the model. He also brought to attention in [46], that any two variables that are independent may not be conditionally independent. Referring back to subsection (2.3.1), if two variables are statistically independent then it means that their joint distribution is the product of their marginal distributions. Therefore what Granger is implying is that variables X and Y may be independent but at the same time variables $X|Y$ and $X|Z$ may not be independent.

Thus one can say that Z brought about the dependency between X and Y. And since we have defined 'causality' as a sort of dependency over a certain causal lag, then one can also expect that there will be cases where Z brings about a causal effect between X and Y. [17]

speaks about latent variables (hidden variables not directly incorporated in the model) that might give rise to correlations in a model where 'causality' is supposed to be inferred from. Dufour [31] mentions indirect causality that might be induced by an auxiliary variable Z on the 'causal' relationship between X and Y. Pearl talks about mediation in his paper [80] where the term direct effect refers to an effect that is not mediated by other variables in the model and by saying this he acknowledges that there exist mediation [16, 83] needed for some (or most) type of causal relationship.

The previous discussions clearly indicates that there is a need to include more than just two variables in a 'causal' analysis. Take lightning and thunder for example [45], we now know that the reason we usually observe lightning before thunder is because light is much faster than sound. We also know the fact that lightning and thunder are both essentially the same event manifested at different times and caused by the same electrical discharge. Let X be thunder, Y be the lightning and Z be the electrical discharge. If we only look at X and Y we will mistakenly say that Y causes X i.e. lightning causes thunder. However if we include Z then we will be able to infer that Z (the electrical discharge) is the real cause of X and Y as well as the fact that the very existence of a relationship between X and Y depends completely on Z.

This kind of 'causality' is what we shall refer to as indirect 'causality' where Z indirectly causes the relationship between X and Y. Whereas the relationship between Z and X as well as the relationship between Z and Y can be said to be a direct cause. By that definition, the electrical discharge directly causes the thunder and it also directly causes the lightning but the electrical discharge indirectly causes the thunder to be related to the lightning. The condition that a causal relation cannot be due to a common cause is referred to as causal sufficiency and some philosophers claim that only direct 'causality' can be considered to be a real 'causality' [50]. Indirect 'causality' is a problem in many fields [45] and we believe that the brain is no exception. To incorporate this indirect 'causality' is very much a problem when the 'causality' measure is not model-free since we have to incorporate all the right variables into the model to begin with. This is one of the main reasons why we will mainly be in favor of model-agnostic 'causality' measures.

3.3 Causality on the brain

Unfortunately, in a complex system where one would expect synchronization and cooperative behaviour, the causal relationship is very complex [52] and the understanding of cause and effect in complex system is definitely lacking [17]. To approximate 'causality' for complex system such as the brain, we first need to have an idea of what we seek from the data sets.

3.3.1 Causal connectivity on the brain

We seek to understand the information transfer and causal connectivity of the brain. The main reason we wanted to establish 'causality' in the brain is to uncover the directed connectivity of the brain. The kind of 'causality' measures we utilize depend on what kind of connectivity we wish to uncover in the brain. In neuroscience, effective connectivity is the term often used for the connectivity that aims to identify the underlying physiological influences of neurons using available time series data. The effective connectivity is defined as the directed influence that a neuronal populations in one brain area exerts on another [41].

Another type of connectivity that does not necessarily require any physiological verification is coined as the dynamical connectivity [17]. The dynamical connectivity is valid when a few issues are taken into account. The first one is the fact that studies have demonstrated [92] that the same physical network structure on the brain can give rise to multiple distinct connectivity depending on interactions with environment. Secondly, neural dynamics is said to alter underlying structural dynamics [17], for example in terms of memory and learning.

Furthermore in our current state of knowledge, knowing all the variables involved with a certain structure will be quite impossible making effective connectivity perpetually provisional (unless perhaps validated by intervention procedures). On the other hand, the dynamical connectivity is a description of dynamical relations between variables regardless. It will be best if one did obtain the effective connectivity where the dynamics and structure go hand in hand and one verifies the other, however in light of the brain as a complex system, this effective connectivity will surely be ever changing. We might want to take things one step at a time and make sure we understand the dynamics first.

3.3.2 Approaches to determining neural causal connectivity

There exist different approaches to achieving causal connections in the brain. One approach to modelling the brain is by utilizing the knowledge of biology and neuroscience to preemptively make the best guess of a model that will fit the brain. Afterwards data sets are fitted to verify the model, this approach is called confirmatory approach [41]. The second approach called the exploratory approach takes the opposite position of inferring the model from the data. This approach does not rely on any preconceived idea and let the data from the brain shape the directed model of the brain. This view of modelling is also taken by other fields and there is a growing general view that biology should move from hypothesis directed research to exploratory methods [16]. Indeed, nature has so many secrets that humans might benefit from putting assumptions aside and listening to it without attaching preconceive notions.

One can think of the different approaches as being on a spectrum from purely confirmatory to purely exploratory. An example of a method that is often classed as being near the confirmatory end of the spectrum is the Direct Causal Model (DCM) introduced by Friston [41] and the graphical model [76]. DCM incorporates explicit model of the neuronal causes and is usually used to infer effective connectivity [17]. One can say that G-causality and Transfer Entropy resides near the other end of the spectrum since both derive inferences directly from data and conclusions are made based on distribution of the sampled data. Henceforth we will focus more on the exploratory end.

However, G-causality is also confirmatory in sense that it assumes autoregressive process. Transfer Entropy seems more exploratory than G-causality from this point of view. Recent implementations of DCM incorporate evidences from data in model selection process thus becoming somewhat exploratory [17]. The two approaches seems to be converging more and more.

3.3.3 Establishing connectivity through EEG

If one intends to pursue 'causality' the exploratory way, EEG or MEG data would be the preferable to the fMRI. This is due to the fact that fMRI data changes with the structural model which implies that one cannot directly compare different regions of the brain without

a certain amount of structural model selection [41]. What we want to do is to get an insight of the inner workings of the brain through a method which does not require direct intervention in the brain by analysing EEG data sets. Wiener being keen on causality and information theory has pointed out how EEG may be useful for this purpose when he [107] wrote:

> "Or again, in the study of brain waves we may be able to obtain electroencephalograms more or less corresponding to electrical activity in different parts of the brain. Here the study of the coefficients of causality running both ways and of their analogs for sets of more than two functions may be useful in determining what part of the brain is driving what other part of the brain in its normal activity".

The normal activity Wiener is referring to here is activity on the brain without any intervention of artificial stimuli which he claims might bring about artifacts.

Artifacts such as movements and eye twitches (manually removed by neuroscientist) are usually an issue when dealing with EEG data sets because it gets in the way of time relation. The bandpass filtering that often has to be done on EEG data sets is also said to be damaging to G-causality estimation [17]. In terms of the data sets we have obtained, due to the use of the best possible equipments supplied by Björn's team, almost no artifact removal is needed and very minimal filtering is required. Therefore we are confident that we have a good set of data to test our results on.

Recall that in our data sets, EEG refers to the recording of the brain's spontaneous electrical activity over a short period of time, as recorded from eight electrodes placed on the scalp. It has been said that although the application of 'causality' measures on EEG data can be extremely useful due to its sub-millisecond time resolution, it also suffers from uncertainties in source space localization [17]. However, if we are focusing on the dynamical connectivity of the brain, this is a question that we can put aside for the moment. Here we assume that each electrode detects an average voltage of its surroundings thus each electrode represents a spatially averaged electrical activity at one point on the skull. We can think of it as the collective activity of neurons in that area of the scalp.

The notion that A causes B if A in the past incites B in the present (or it's relative future) is what we will define as 'causality' in our context. And this is what we will be looking for in the brain. In terms of EEG electrodes, if a certain electrode A in the past consistently incites a certain electrode B in the present, then we shall say that electrode A causes electrode B and this then translates to the area of the scalp. The general idea is that if electrode A causes electrode B we would want to be able to detect it.

3.4 Granger Causality

First introduced by nobel prize winning Clive Granger [44] in the context of linear autoregressive (AR) model, G-causality is the most commonly used 'causality' indicator [17]. In his paper Granger explains the direction of 'causality' in a simple two-variable (binary) model. G-causality is also often known as the Granger-Wiener causality especially since Granger himself quoted that Wiener inspired him in his nobel lecture [52]. Granger outlined two things about 'causality' in that same lecture. The first is that the cause must come before the effect (the arrow of time) and secondly that the cause should contain unique information about the effect that cannot be found in any other variable [46].

3.4.1 G-causality: An overview

Referring back to our three steps in ascertaining whether variable Y causes variable X. We shall go through these steps again with a Granger causality point of view. The first step was to predict the current value of X using the historical values of X. The second step is to do another prediction where the historical values of Y and X are both used to predict the current value of X. In order to accomplish the first and second step, we need to fit variables X and Y into a model.

It is worth pointing out that, this is not a trivial task. First and foremost to fit variables [68, 52] X and Y into a model (usually some form of AR process) requires some kind of method (usually standard linear regression method). Amongst the more popular ones are the least square method or the Yule method [17]. For example one could use values of $(X)_1$ to $(X)_n$ to predict $(X)_{n+1}$ and for the second prediction one could utilize $(Y)_1$ to $(Y)_n$ as

well as $(X)_1$ to $(X)_n$ to predict $(X)_{n+1}$. However, how long a history should be taken into account for prediction also needs to be decided upon. This process is usually referred to as determining the order of the model and it requires fitting to the data sets. The order (model fit) is determined using certain criterions for example Bayesian Information Criterion and Akaike Information Criterion [17]. Only after satisfying ourselves that the model with a certain order is a good fit, does one proceed into predicting the values.

After the first and second prediction, the third step is comparing the former to the latter i.e. ascertaining whether the second prediction is any better than the first. In order to do this, statistical significance is required and test statistics are used, amongst them the Granger-Sargent [52, 17] and the Granger-Wald test [52]. And if the chosen test is satisfied, one can conclude that Y Granger causes X which is often written as Y G-causes X.

In short, one can say that G-causality works on the premise of comparing predictions based on linear regression. The variables need to be linearly regressed in order to get linear equations in an AR model. The equations will be utilized in the form of two predictions. The first to predict X using its history and the second predict X again using the history of both X and Y. If the second prediction is deemed to be better than the first, we can happily say that Y G-causes X.

3.4.2 Challenges to G-causality

There are a few issues to focus on here, first and foremost the very fact that we need to linearly regress the data to obtain the prediction means that we will loose a lot of nonlinear information. The usual argument of the proponents of G-causality is that linear approximation works well on large scale interactions [17]. However, to Granger's credit [44, 45, 46] he has always been clear that G-causality is not absolute 'causality' and he himself acknowledges that the optimal predictor may very well be nonlinear [44]. It has been pointed out by many [15, 17, 68] that due to obvious nonlinear dependencies on neuro data sets, using G-causality may not suitable.

This brings us to an essential point which is the modelling itself. Granger concedes that it is not an easy task to get the modelling right and missing variables may lead to spurious values [44]. The fitting of the linear regression to the data sets, the order of the model and

the determination of which predictor is better are all determined by some form of criterion that has to be decided upon and serves in itself as a varying parameter. The order of the model is always a problem, too low an estimation leads to a poor representation of the data and too high an estimation could also lead to various problems [17].

Moreover, G-causality was designed as a pairwise measure and is not suitable to deal with three variables or more [10]. This is important since not only do we want to incorporate general dependencies but we also want to be able to eventually work with more than two variables. Latent variables (or indirect causality) have been known to cause spurious causal interaction when using G-causality [17].

3.4.3 Generalization and extensions of G-causality

The need to include more variables and nonlinear elements in examining data sets has resulted in many new forms of generalized G-causality. Generally these extensions fall into a few categories. The first category includes attempts to extend G-causality to be able to include more than two variables. One example of this, is called the multivariate Granger causality (MVGC) [10, 17] is also an extension of another variant of G-causality called conditional Granger causality. The MVGC utilizes the determinant of the residual covariance (the generalized variance).

Another slightly different extension utilizes total variance which is the trace of the residual covariance matrix instead of the determinant [10, 17]. MVGC is said to be superior to G-causality since it is not only able to quantify more than two variables but it can even quantify interaction between groups of variables. Moreover MVGC has been proven to be equivalent to Transfer Entropy on a Gaussian distribution [11, 10]. To address the problem of indirect causality, Partial G-causality was introduced. It is said to be able to mitigate the influence of latent variables [17].

The second category of extensions aim to extend the definition of G-causality to include nonlinearity. One example is the attempt to extend the definition of G-causality to nonlinear bivariate time series by utilizing nonlinear radial basis functions [68, 17]. Another idea for incorporating nonlinearity involves locally linear models. The idea is to divide the data into local neighborhoods where it will be approximated by a linear model and G-

causality is applied. The extended Granger causality indices is given by averaging over the neighborhood sampling [17, 68].

Another category of extensions can be classed as the nonparametric approach to generalizing G-causality. The most pressing issue with G-causality so far is that model must be matched to underlying dynamics and may lead to spurious values. Nonparametric approaches are intended to be model free. Most of these types are rooted in information theory. Generalized correlation integrals and conditional entropy among them [17]. The Transfer Entropy is also said to be the information theoretic approach to G-causality [52]. This is because of the similarities they have in concept and approach to quantifying 'causality'.

Chapter Summary

The notorious question of 'causality' has long been debated. The definition of it is yet to be agreed upon and can lead to intense philosophical debate. In this chapter, we have defined 'causality' as being prediction based and we have addressed some issues related to it. We then went on to consider the brain itself and the differing views of how to understand its causal relationships. We discussed how we decided to focus more on the exploratory end of the modelling spectrum to establish dynamical rather than effective connectivity. G-causality was considered given its popularity, thus some challenges and extensions of G-causality was discussed. One of the information theoretic extension of G-causality is said to be the Transfer Entropy. It appears that we have come full circle, arriving at Transfer Entropy from both the nonlinearity and the 'causality' end. It is now high time we proceed to look into this measure with more clarity.

Chapter 4

Transfer Entropy

From previous chapters it seems apparent that the Transfer Entropy could be key in achieving our aims of capturing nonlinearity and ‘causality’ in the brain. We first explain the idea behind this measure and how the Markov property plays a key role in the definition. We highlight the difference between Transfer Entropy and conditional Mutual Information in terms of transition and marginal probability. We then look at the definition of Transfer Entropy from a prediction point of view and contrast it with G-causality. We point out the weaknesses and strong points of this measure. Subsequently, we go through some Transfer Entropy literature and examine more of the challenges it faces. Last but most certainly not least, the simplest case is revisited and Transfer Entropy for causal lag detection is highlighted.

4.1 Transfer Entropy and the Markov property

Recall the definition of Transfer Entropy in equation (2.27), where it was introduced as a conditional Mutual Information variant with incorporated time delays. Schreiber’s original definition of the Transfer Entropy however was based on the generalized Markov Property and transition probabilities [10, 11, 57, 89].

4.1.1 The transition probability

In [89], Schreiber points out that in order to incorporate the dynamical structures, transition probability should be used instead of simultaneous (static) probability. To illustrate the difference, imagine a ball that can be red or blue at any given time step and let the process run for 10 time steps. If the ball was red 6 out of 10 times, counting the frequencies yield that the static probabilities of the ball being red (R) and blue (B) is, $P(R) = 0.6$ and $P(B) = 0.4$ respectively.

The transition probabilities however, take into consideration the order of change. If the ball was red six times before it was blue i.e RRRRRRBBBB then the probability for the ball to stay red is, $P(R \rightarrow R) = \frac{5}{9}$ and the probability for it to change from red to blue is $P(R \rightarrow B) = \frac{1}{9}$. Similarly $P(B \rightarrow B) = \frac{3}{9}$ and $P(B \rightarrow R) = 0$. While the static probabilities remain the same for the case where the order of change is RBRBRBRBRB, the transition probabilities do not. Now, we have that $P(R \rightarrow B) = \frac{5}{9}$, $P(B \rightarrow R) = \frac{4}{9}$, $P(B \rightarrow B) = 0$ and $P(R \rightarrow R) = 0$. Transition probabilities of higher order (more than one time step) can also be defined [77].

Basically, the transition probability considers the state of the ball (red or blue) at different time steps and the changes it makes, instead of calculating the frequencies of the ball being red or blue throughout the time steps. The transition probabilities, capture the dynamics in this sense. The essential difference between conditional Mutual Information and Transfer Entropy is that the latter utilizes transition probabilities in place of static probability. It is worth pointing out again that in order to get some sort of directionality or 'causality', measuring values across different time steps is somewhat inevitable.

A Markov process (also known as a Markov chain) is a random process that retains no memory of where it has been in the past [77]. Therefore, the state of the system in the future only depends on the present. If X is Markov process [29] with possible values $x_i, i = 1 \cdots n$, then we have that

$$P(X_{n+1} = x_{n+1} | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

This memoryless property is called the Markov property. In [89], a system of Markov process of order k was considered. A Markov process of order k is a random process that

retains memory of only k steps in the past so that

$$P(X_{n+1} = x_{n+1}|X_1 = x_1, \cdots, X_{n-k-1} = x_{n-k-1}, X_{n-k} = x_{n-k}, \cdots, X_n = x_n) \quad (4.1)$$
$$= P(X_{n+1} = x_{n+1}|X_{n-k} = x_{n-k}, \cdots, X_n = x_n).$$

Let the $X_n^{(k)} = (X_{n-k} = x_{n-k}, ..., X_n = x_n)$, so that the Markov property in equation (4.1) can be written as $P(X_{n+1} = x_{n+1}|X_n^{(n-1)}) = P(X_{n+1} = x_{n+1}|X_n^{(k)})$.

A generalization to this property [89] is when we include another variable in the condition $P(X_{n+1} = x_{n+1}|X_n^{(n-1)}, Y_n^{(n-1)}) = P(X_{n+1} = x_{n+1}|X_n^{(k)}, Y_n^{(l)})$, where Y is a Markov process of order l. This implies that variable X depends on the history of variable Y up to order l. The idea of Transfer Entropy incorporates the generalized Markov property in determining whether there is a flow of information from one process to another.

4.1.2 Schreiber's Transfer Entropy

Noticing that Mutual Information and the other entropy based measures mentioned before, did not capture directional information and the dynamics, Schreiber [89] introduced the Transfer Entropy. The Transfer Entropy of Y to X, $T_{Y\to X}$ [11, 52, 57, 89] (where X is a Markov process of order k and Y is a Markov process of order l) is given by

$$T_{Y\to X} = E\left[log\, \frac{P(X_{n+1} = x_{n+1}|Y_n^{(l)}, X_n^{(k)})}{P(X_{n+1} = x_{n+1}|X_n^{(k)})} \right] \quad (4.2)$$
$$= \sum_{x_{n+1}} \sum_{x_n} \sum_{y_n} P(X_{n+1} = x_{n+1}, Y_n^{(l)}, X_n^{(k)})\, log\, \frac{P(X_{n+1} = x_{n+1}|Y_n^{(l)}, X_n^{(k)})}{P(X_{n+1} = x_{n+1}|X_n^{(k)})}.$$

We again take the value $0\, log\, 0 = 0$. If $T_{Y\to X} \neq 0$ and $T_{X\to Y} = 0$ then one can say that Y 'causes' X [89, 57]. The Transfer Entropy can be identified as a variant of conditional Mutual Information. Recall the definition of Transfer Entropy in equation (2.27), where Transfer Entropy was introduced as a version of time delayed conditional Mutual Information such that $I(X^1, Y|X) = H(X^1|X) - H(X^1|Y, X)$. Here X^1 is the variable X shifted one time step so that the values of X^1 is always one time step ahead of X. This is a simple

case of Schreiber's definition of the Transfer Entropy when $k = 1$ and $l = 1$ such that

$$T_{Y\to X} = E\left[log\, \frac{P(X_{n+1} = x_{n+1}|Y_n = y_n, X_n = x_n)}{P(X_{n+1} = x_{n+1}|X_n = x_n)} \right] = E\left[log\, \frac{P(X^1|Y, X)}{P(X^1|X)} \right] \quad (4.3)$$
$$= H(X^1|X) - H(X^1|Y, X) = I(X^1, Y|X).$$

Clearly it is the transition probabilities $P(X_{n+1} = x_{n+1}|Y_n = y_n, X_n = x_n)$ as well as $P(X_{n+1} = x_{n+1}|X_n = x_n)$ that are taken into account. Contrast this with the definition of conditional Mutual Information $I(X, Y|Z)$ in equation (2.15) where $P(X_n = x_n|Y_n = y_n, Z_n = z_n)$ and $P(X_n = x_n|Y_n = y_n)$ are considered instead.

In the original paper, Transfer Entropy was intended to measure the deviation from Markov property. Schreiber's aim was to incorporate the properties of Mutual Information and the dynamics captured by transition probabilities in order to understand the concept and exchange of information. Taking into account two processes at different time steps comes about naturally as soon as transition probabilities are considered. Both Transfer Entropy and time delayed Mutual Information were defined to incorporate time delay, however time delayed Mutual Information does not utilize transition probabilities. Recall the definition of time delayed Mutual Information such that $I(X, Y^\tau) = E_{p_{XY}(x,y)}\left[log \frac{P(X,Y^\tau)}{P(X)P(Y^\tau)} \right]$. The probabilities that are taken into account are the joint probability $P(X, Y^\tau)$ and the marginal probabilities $P(X)$ as well as $P(Y^\tau)$.

Schreiber in [89, 57] claims that time delayed Mutual Information fails to distinguish common history of the stochastic process. To prove his point he experimented on spatiotemporal systems with no coupling where both time delayed Mutual Information and Transfer Entropy gave zero values. However when coupling was present [89, 75], the time delayed Mutual Information reflected static (as opposed to dynamical) properties and gave nonzero values in both directions whereas Transfer Entropy was nonzero for one direction only and hence indicating clear directionality. This is also the findings of [40], which addresses indirect 'causality' within the time lags of a certain process say X. The Transfer Entropy on itself $T_{X\to X}$ is compared to the auto Mutual Information $I(X, X^\tau)$ on a coupled Lorentz system where the coupling is controlled. The results clearly indicate that while auto Mutual Information could not differentiate direct and indirect 'causes' within

the different time steps, the Transfer Entropy correctly detected only the direct ‘causes’.

4.1.3 Directionality of couplings in dynamical systems

As mentioned before, many approaches of inferring ‘causality’ came about due to investigations on synchronization of chaotic systems. In the introduction paper [89], Transfer Entropy was introduced as a measure to detect the direction of couplings in dynamical systems when it was tested on a unidirectionally coupled maps and the Ulam map before application on real data sets. In [57], the Transfer Entropy is used to study the information content in relation to synchronization. Similarly Transfer Entropy or some closely related variant of conditional Mutual Information was tested on logistic maps [82], Ulam maps [68], Hènon map [68], Lorentz systems [40, 82], Rössler models [101, 78], Ornstein-Uhlenbeck process [86] and various other forms dynamical systems. In [78, 101, 40, 82], the conditional Mutual Information is applied directly on the generated values of these dynamical systems. However, in [78] the phase of the coupled oscillator in used instead of the actual values. In all these papers the results from using these variants of Transfer Entropy on the dynamical systems were dominantly positive.

In [68], the performance of a few different methods for testing ‘causality’ were evaluated on various forms of Ulam maps and Hènon maps. This was done in order to assess the usefulness of these methods for detecting asymmetric couplings and directional of information flow in a deterministic chaotic system. Among the methods tested include Transfer Entropy and some extensions of G-causality. The conclusion of the paper was that their first choice given a priori unknown dynamics will be Transfer Entropy. On a more theoretical end of the spectrum, [66] attempts to rigorously formalize information transfer between dynamical system components for systems with fully known dynamics. The transfer of entropy (the amount of entropy transferred between processes) is the focus this paper. It was remarked in [66] that the findings are consistent with Schreiber’s Transfer Entropy. The results were applied on systems with Hènon map and Baker transformation.

4.2 Transfer Entropy and G-causality

Schreiber [89] defined Transfer Entropy to measure deviation from the generalized Markov property. One can also look at Transfer Entropy as a generalization of G-causality [52] as well as a 'causality' measure that is inline with G-causality in terms utilizing prediction as a means to infer 'causality' [11]. Indeed there are papers claiming that Transfer Entropy equals G-causality when the distribution is Gaussian [11, 10].

4.2.1 Transfer Entropy as a method that compares predictions

Transfer Entropy can be compared to Granger Causality step by step as a prediction method. Say we have processes (variables) X and Y. We wish to test whether Y causes X. To visualize Transfer Entropy we shall refer to the $T_{Y\to X} = E\left[log \frac{P(X_{n+1}=x_{n+1}|Y_n^{(l)},X_n^{(k)})}{P(X_{n+1}=x_{n+1}|X_n^{(k)})}\right]$ definition. Referring back to the three steps we have outlined for ascertaining 'causality', the first step would be to predict the current value of X using the historical values of X. In Transfer Entropy this will be the process of obtaining the value of transition probability $P(X_{n+1} = x_{n+1}|X_n^{(k)})$. Contrasting this to the first step when using G-causality, at this stage we should first have a model (usually Auto Regressive) that can incorporate both variables X and Y. We have pointed out in subsection (3.4.1) that finding the right order of the model is an issue for G-causality application. This is also true for Transfer Entropy which requires the denominator $P(X_{n+1} = x_{n+1}|X_n^{(k)})$ to be estimated for the prediction. The order comes in the form of deciding what the value of k is. By assuming kth order of the Markov property, one is incorporating k historical values into the prediction.

Second step, another prediction where the historical values of are Y and X are both used to predict the current value of X. For Transfer Entropy this requires the estimation of $P(X_{n+1} = x_{n+1}|Y_n^{(l)}, X_n^{(k)})$. The same issue of determining the value of l in relation to Y needs to be taken into account. G-causality makes use of the preconceived model to do its prediction at this second step. The last step is to compare the former prediction to the latter. If the second prediction is judged to be better than the first, one can say that Y causes X. Transfer Entropy utilizes the expected log ratio between the two probability distributions to compare the predictions hence $T_{Y\to X} = E\left[log \frac{P(X_{n+1}=x_{n+1}|Y_n^{(l)},X_n^{(k)})}{P(X_{n+1}=x_{n+1}|X_n^{(k)})}\right]$. G-causality mainly compares the variance of the error terms of both predictions in the model

to determine which one is better. In order to do this for G-causality test statistic are used [52, 17, 11]. According to Schreiber's definition [89, 57] for Transfer Entropy, only if $T_{X\to Y} = 0$ and $T_{Y\to X} \neq 0$ can we conclude that Y causes X. In later papers it has been suggested that significant tests in the form of surrogates [101, 102, 78, 75] could be sufficient.

Transfer Entropy and G-causality was proven to be equivalent on the Gaussian model [11, 10]. This is due to the fact that Gaussian distributions can be calculated analytically from the covariance matrix Σ. The definition of entropy given a Gaussian distribution is

$$H(X) = \frac{1}{2}\, ln(|\Sigma(X)|) - \frac{1}{2}\, ln(2\pi e)$$

where $|\Sigma(X)|$ is determinant of the covariance matrix of X. One can also obtain the residuals of a linear regression in terms of this covariance matrix therefore enabling one to link the G-causality and Transfer Entropy analytically [57]. It appears that under the Gaussian assumption there is nothing additional to account for the nonlinear extensions of G-causality since the Gaussian AR process is necessarily linear [11].

4.2.2 Transfer Entropy versus G-causality

Other than being model agnostic and nonlinear, Transfer Entropy also easily extends to multivariate applications. One example of this is the generalization of Mutual Information $MI(X,Y,Z) = H(X)+H(Y)+H(Z)-H(X,Y,Z)$ in subsection (2.4.2). More than one variable can be incorporated as X, Y or Z since the only thing that matters is the probability distribution. In fact the multivariate concept of Transfer Entropy together with entropies and Mutual Information have been suggested as unifying frameworks for determining directed networks [26, 86]. However, the estimation of the probability distribution remains a big obstacle to successfully implementing this framework. Non-information theoretic based extension of G-causality to multivariate concept is not generally straightforward as explained in subsection (3.4.3) and the estimations of many variable brings about similar estimation challenges.

Being model free, Transfer Entropy is usually preferred for the exploratory approach, however not having an underlying model is not always advantageous, the high sensitivity

and the uncertainty of how to interpret the outcome may be a problem [102]. The main strength of G-causality is that it is well defined and pragmatic as a result of being applied in a well understood framework. Although one does not expect this to be the case where the nonlinearities of a complex systems is considered (EEG data sets have been said to be not even approximately Gaussian [57]), it is well known that if the interactions are indeed linear, linear methods such as G-casuality will usually outperform Transfer Entropy [17, 102]. Moreover it was pointed out in [11] that even though for most empirical data it is difficult to establish the extent to which Gaussian assumptions are tenable, it is nonetheless widely employed. If this is indeed the case (such that Gaussian distributions apply) then the two methods are interchangeable.

Furthermore, G-causality is simpler to deal with because the sample statistic is known and therefore there exist many forms of significant test to choose from when comparing the predictions (in the last step). For Transfer Entropy it has been said that a significant test would be hard to devise due to the unknown sample statistic [11]. However as we pointed out before, in current applications of Transfer Entropy, surrogates are used as a form of significant testing [101, 102, 78, 75].

4.3 Challenges to Transfer Entropy

The main challenge to applying Transfer Entropy would have to be the estimation on real data sets due to probability estimation. However, we put aside this issue for a while and we will return to this issue in later chapters before the treatment on real data sets. The challenges addressed in this section apply to both Transfer Entropy and G-causality. Essentially, these are the challenges that are usually associated with prediction based 'causality' measures.

4.3.1 In addressing deterministic cases and full synchronization

At the very beginning, Schreiber [89] pointed out that Transfer Entropy was meant for cases where neither of the systems nor their couplings may be assumed to be deterministic. Consequently, if Y is completely determined by X, then $T_{Y\to X} = 0$. This will be the case

for deterministic coupling since the probabilities of Y would be exactly the same as X. Thus Transfer Entropy with itself would be

$$T_{X\to X} = E\left[log\,\frac{P(X_{n+1}=x_{n+1}|X_n^{(k)},X_n^{(k)})}{P(X_{n+1}=x_{n+1}|X_n^{(k)})}\right]$$
$$= E\left[log\,\frac{P(X_{n+1}=x_{n+1}|X_n^{(k)})}{P(X_{n+1}=x_{n+1}|X_n^{(k)})}\right] = log\,1 = 0.$$

This constitutes the reason why Mutual Information and Transfer Entropy is invariant under diffeomorphic (isomorphic on smooth manifolds) transformations [57]. On one hand, it makes sense that if the relationship between X and Y makes them practically indistinguishable (from the equation above we can see this is due to the probabilities) from each other, there should be no information flow.

However one could argue that, instead of having no flow of information at all, that this is the case where there is a complete flow of information from one process to another, or in other words that it is fully synchronized. Indeed, not only is Transfer Entropy zero for deterministic cases but it is also zero for complete synchronization [57, 102]. In a way, this is a bit worrying since on one hand $T_{Y\to X} = 0$ implies that Y is completely independent of X, but as we have just seen $T_{Y\to X} = 0$ could also imply that Y is completely dependent on X (deterministically coupled or fully synchronized). This issue has been pointed out by [76] which claims that Transfer Entropy does not coincide with information flow and suggests a new measure in the causal Bayesian network to overcome this.

This same issue is touched upon in [40] where it is said that conditional Mutual Information will only work if the underlying processes has a varying source of entropy (stochastic or chaotic) and thus one process is not function of the other. This is the kind of information that is addressed by Transfer Entropy where the reduction in uncertainty is taken as information transfer. Hence, according to Transfer Entropy there is no information flow when the variables are indistinguishable and at complete synchronization. Moreover, one could say that a causal direction is impossible to establish in this case (refer to discussion in subsection (3.2.2)).

4.3.2 Indirect 'causality'

We have touched upon the issue of indirect 'causality' in subsection (3.2.3). There is a real need to differentiate instances of direct and indirect dependencies. If there was only three stochastic processes X, Y and Z in a system, then theoretically $I(X,Y) = I(X,Y|Z) > 0$ implies that X is directly dependent on Y and vice versa. One could say that the link between X and Y is direct and does not depend on any other variable. However if we have that $I(X,Y) \neq I(X,Y|Z)$, then one can say that the dependency of X on Y (or Y on X) is indirect and depends on Z. In other words, the interdependency between X and Y indirectly depends on Z.

In order to incorporate time and direction into this idea, [40] has made used of this idea with values of $Y = X^{\tau_1}$ and $Z = X^{\tau_2}$ where X^{τ_1} and X^{τ_2} are both time shifted versions of time lag τ_1 and τ_2 respectively, so that $I(X, X^{\tau_1})$ and $I(X, X^{\tau_1}|X^{\tau_2})$ can be compared on a Lorentz systems where the coupling can be controlled. The auto Mutual Information $I(X, X^{\tau_1})$ could detect the dependencies over time but could not differentiate between direct and indirect ones. Whereas the conditional Mutual Information (or the Transfer Entropy value used on a single variable) $I(X, X^{\tau_1}|X^{\tau_2})$ gives clear indications of which time lags is directly 'causing' the others. In [40], the conclusion was that the conditional Mutual Information is able to reveal direct and indirect causality.

The idea is that one can always condition out other variables (be it time shifted or not) by incorporating the variables in Z (which can be multivariate) in the term $I(X,Y|Z)$ (which also applies to Transfer Entropy) and say that this represents the direct interdependency of X and Y without the effects of Z. However, Schreiber [89] has forewarned that conditioning on too many variables is dangerous as the estimations will be much more difficult. To overcome this problem several alternative solutions have been proposed [86, 82, 31].

4.4 Incorporating time delays

The idea of indirect 'causality' is often related to the existence of time delay between cause and effect. In subsection (3.2.3), the fact that causal lags will inevitably exist for

'causality' to be manifested was established. The example of the lightning and thunder was highlighted where lightning would appear to be causing thunder if the real cause of electrical discharge was not taken into account. In a way one could say that lightning indirectly causes thunder, on the other hand one could also say that lightning and thunder are the same events manifested at different time lags. In other words, thunder can be considered as a delayed effect of lightning. Thus, identifying time delay is a big step towards identifying indirect 'causality'.

4.4.1 The detection of 'causal' lags

The 'causal' lag is the time delay that exist between the 'cause' and effect. It could be interpreted as the time taken to deliver the information from the causal variable to the affected one. We suspect that these types of delays should be present in the brain where neurons are constantly firing. Therefore the past and future (in terms of the lags) must be at a rate that is meaningful and this is where EEG with its high temporal resolution should be most helpful.

This was exactly the point made out by [102] which claims that in neuroscience, the interaction may involve large time delays of unknown duration. Therefore [102] recommends that a time shift test in taken addition to Transfer Entropy whenever multiple source signal is likely to be present especially in EEG data. The time shift test proposed was simply looking at various Transfer Entropy values for different time shifted version of a certain variable (process). In particular, what [102] implemented was that if two variables X and Y has Transfer Entropy values that indicates Y causes X, let this be the hypothesis. $T_{Y^1 \to X}$ is calculated where Y^1 is the time shifted variable Y such that $Y_t^1 = Y_{t+1}$. If $T_{Y^1 \to X} \geq T_{Y \to X}$, then it is concluded that the relationship is due to instantaneous mixing and the idea that Y causes X is discarded. Otherwise the hypothesis is accepted. The usage of Transfer Entropy in combination with time shift test was recommended on EEG data.

4.4.2 Simplest case and 'causal' lag detection

The simplest case of equation (4.2) is when we let $l = 1$ and $k = 1$ so that

$$\begin{aligned} T_{Y\to X} &= E\left[log\, \frac{P(X_{n+1} = x_{n+1}|Y_n^{(1)}, X_n^{(1)})}{P(X_{n+1} = x_{n+1}|X_n^{(1)})} \right] \\ &= E\left[log\, \frac{P(X_{n+1} = x_{n+1}|Y_n = y_n, X_n = x_n)}{P(X_{n+1} = x_{n+1}|X_n = x_n)} \right] \\ &= E\left[log\, \frac{P(X_n = x_n|Y_{n-1} = y_{n-1}, X_{n-1} = x_{n-1})}{P(X_n = x_n|X_{n-1} = x_{n-1})} \right]. \end{aligned}$$

We seek to investigate this form especially since Schreiber himself warns against conditioning on too many variables. Notice that the calculation will be just as simple if we defined

$$\begin{aligned} T_{YX}^{(\tau)} =& E\left[log\, \frac{P(X_n = x_n|X_{n-1} = x_{n-1}, Y_{n-\tau} = y_{n-\tau})}{P(X_n = x_n|X_{n-1} = x_{n-1})} \right] \\ =& \sum_{x_{n+1}\in\mathcal{X}} \sum_{x_n\in\mathcal{X}} \sum_{y_n\in\mathcal{Y}} P(X_{n+1} = x_{n+1}, X_{n-1} = x_{n-1}, Y_{n-\tau} = y_{n-\tau})\times \\ & log\, \frac{P(X_{n+1} = x_{n+1}|X_{n-1} = x_{n-1}, Y_{n-\tau} = y_{n-\tau})}{P(X_{n+1} = x_{n+1}|X_{n-1} = x_{n-1})} \end{aligned} \tag{4.4}$$

where $\mathcal{X}$ is the state space of X and $\mathcal{Y}$ is the state space of Y. We take the value $0\, log\, 0 = 0$. Writing this equation in terms of shifted variables X^{-1} and $Y^{-\tau}$ we have that

$$\begin{aligned} T_{YX}^{(\tau)} =& E\left[log\, \frac{P(X_n = x_n|X_{n-1} = x_{n-1}, Y_{n-\tau} = y_{n-\tau})}{P(X_n = x_n|X_{n-1} = x_{n-1})} \right] = E\left[log\, \frac{P(X|X^{-1}, Y^{-\tau})}{P(X|X^{-1})} \right] \\ =& \sum_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{XX^{-1}Y^{-\tau}}(x, x', y)\, log\, \frac{p_{X|X^{-1}Y^{-\tau}}(x|x', y)}{p_{X|X^{-1}}(x|x')}. \end{aligned} \tag{4.5}$$

In relation to equation (2.27), we highlight that the formula of the Transfer Entropy in equation (4.5) is related to conditional entropy and conditional Mutual Information in this

way,

$$\begin{aligned} T_{YX}^{(\tau)} &= E\left[log\,\frac{P(X|X^{-1},Y^{-\tau})}{P(X|X^{-1})}\right] = E\left[log\,\frac{P(X,X^{-1},Y^{-\tau})}{P(X^{-1},Y^{-\tau})P(X|X^{-1})}\right] \\ &= E\left[log\,\frac{P(X,Y^{-\tau}|X^{-1})}{P(Y^{-\tau}|X^{-1})P(X|X^{-1})}\right] = I(X,Y^{-\tau}|X^{-1}) \\ &= H(X|X^{-1}) - H(X|Y^{-\tau},X^{-1}). \end{aligned} \tag{4.6}$$

This simple form allows us to vary the values of τ in investigating whether there is a certain causal lag required in order to manifest the dependency. This was the form suggested in the time shift test of [102]. This was also the form suggested in [75] where it was called Transfer Entropy. In [75] time delayed Mutual Information was compared to time delayed Transfer Entropy and the conclusion was in favour of Transfer Entropy. A similar idea of causal lag detection called horizons on G-causality is discussed in [31]. Several other approaches to tackling the issue by utilizing permutation entropy are proposed in [65, 82]. If we were to reformulate the time shift test previously discussed, the test will be that if $T_{YX}^{(2)} > T_{YX}^{(1)}$ then the idea that Y cause X should be discarded. In this thesis, we do not necessarily agree with this. We intend to show this situation can occur when Y causes X but not detected at the exact time lag. In fact we shall show that $T_{YX}^{(\tau)}$ for our purposes will be largest at exact causal lag τ.

Chapter Summary

Despite all the challenges facing Transfer Entropy, the usage of the measure has made it one of the most prominent measures in capturing 'causality', often mentioned together or as an extension of G-causality. Even when new measures are proposed [76, 65, 82], the Transfer Entropy becomes the benchmark measure for comparison. This is especially true in the field of neuroscience [17, 102, 5] and in particular when examining EEG data sets [70, 69, 99]. In fact given unknown dynamics, Transfer Entropy was crowned as the first choice among methods for quantifying causal structure of bivariate time series by [68].

Therefore, the Transfer Entropy seems ideal for our purposes. It is a variant of conditional Mutual Information thus non-linear and general when it comes to independence. At

the same time it incorporates the time element that enables ‘causality’ detection. We have taken the simplest Transfer Entropy case and redefined it for causal lag detection. Now we shall put this measure to the test. We have seen that most of the testing on conditional Mutual Information and Transfer Entropy were done on dynamical systems. We aim to look at these values more from a statistical mechanics point of view and we think there is no better testing ground to start with than the famous Ising model.

Chapter 5

The Ising model

The Ising model is a simple mathematical model used in statistical mechanics. Its simplicity makes the two dimensional case exactly solvable [19, 27]. In the first part of the chapter we will explain the concept and some history of the Ising model as well as how it achieves criticality. Some theoretical values of the measures we wish to investigate on the model are introduced. Afterwards, we go on simulating the model using the Metropolis Monte Carlo algorithm. Consequently, we outline the simulation results. Firstly we verify crossover values and a further discussion of what this values implies. Then, we proceed to take a closer look at the Mutual Information, conditional Mutual Information, time delayed Mutual Information and Transfer Entropy applied on the simulated data and what they imply. Lastly, we discuss how Mutual Information and covariance can be related on the Ising model and what this means in relation to dependence of sites on each other.

5.1 Concept of the Ising model

The quest of seeking to explain the macroscopic behaviour of a system on the basis of its microscopic structure in statistical mechanics has its root in the analysis of simplified mathematical models [42]. The Ising model is the simplest of these models. More importantly, phase transition is manifested on the model where a small change in temperature causes a huge change in long range correlative behaviour [27, 28].

5.1.1 About the Ising model

The Ising model was introduced as a simplified representation of intermolecular forces on ferromagnetic metal. This is due to the fact that ferromagnetic metal can be regarded as being composed of elementary magnetic moments called spins which are arranged on the vertices (sites) of a crystal lattice. The phase transition on this lattice is said to be the spontaneous emergence of magnetization in zero external magnetic field as temperature is lowered below a certain critical temperature [27].

The Ising model got its name from the German physicist Ernst Ising who wrote his doctoral thesis on this particular model in 1925 [28, 19] where he utilized the model in trying to explain certain empirically observed facts about ferromagnetic materials. He was a student of Wilhelm Lenz who had earlier roughly proposed the idea in 1920. At first, even Ernst Ising himself gave up research in physics after thinking that he had proven that his model had no physical usefulness [19]. It physically appeared that an oversimplified model representation of intermolecular forces on which this model is based on would make it unapplicable to any real system.

It was only 20 years later that Ising found out that he was famous for other peoples work on his abandoned model [28]. Although his work on the one dimensional Ising model did not achieve phase transition, the two dimensional model does. In fact, an analytical solution has been given by the nobel prize winning Onsager [27]. Currently, the model has been applied to biology, sociology and economics (just to name a few) [27, 28]. Practically, any case where you have two possible states of interacting components to take into account and where cooperative behavior is studied, some form of Ising model can be applied. Indeed, the importance of the Ising model cannot be overstated.

The model may be summarized as follows. Assuming that the physical system can be represented by a regular lattice arrangement and that the sites (particles) are positioned at points of some lattice embedded in Euclidean space. Each site may either be in two states, representing the physical state of spin-up and spin-down. The orientation of each spin is random but subject to spin-spin interaction which favours their alignment. Spin values are chosen at random according to a certain probability measure, known as Boltzmann distribution or Gibbs measure, which is governed by interactions between neighbouring particles.

The theory of Gibbs measure is a branch of classical statistical physics which also can be viewed as part of probability theory. It was proposed as a natural mathematical description of an equilibrium state of a physical system which consists of a very large number of interacting components (such as spins on ferromagnets) [42]. In probabilistic terms, it is none other than the distribution of a countably infinite family of random variables which admit some prescribed conditional probabilities. This notion has received considerable interest from both mathematical physicist and probabilists. The Gibbs measure has been proven to be the unique measure that maximizes entropy and underlies the maximum entropy method.

5.1.2 The mathematical formulation

One can visualize the Ising model as a two dimensional square lattice with length L composed of $N = L^2$ sites (vertices) $s_i, i \in \mathcal{N} = \{1 \cdots N\}$. These sites can only be in two possible states, spin-up ($s_i = 1$) or spin-down ($s_i = -1$). The full description of a microstate or a configuration will be denoted by $\omega = \{s_1 \cdots s_N\}$. Let $(s_i)_\omega$ be the number which appears as the ith component in ω. This number represents the state of the ith site in configuration (microstate) ω. We shall take the liberty of using s_i instead of $(s_i)_\omega$ whenever the configuration ω is understood from the context. Let Λ be the set of all possible configurations or microstates such that $\omega \in \Lambda$.

The interactions between these sites are given by the interaction strength. In this thesis, we restrict the interaction of the sites to only its nearest neighbour (in two dimensions this will be sites to the north, south, east and west). Let the interaction strength between i and j be denoted by

$$J_{ij} = \begin{cases} J \geq 0, & \text{if } i \text{ and } j \text{ are nearest neighbours and } i, j \in \mathcal{N} \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

The nearest neighbour restriction will shape the Ising model to be Markovian in a sense that the probability of a given site $i \in \mathcal{N}$ being in state α is given by

$$P(s_i = \alpha | s_j, i \neq j) = P(s_i = \alpha | s_j, j \text{ is nearest neighbour of } i),$$

where instead of being dependent on all other sites on the lattice it is only dependent on the nearest neighbour. This property is sometimes referred to as the Markov random field [42]. The Hamiltonian (energy), $\mathcal{H}$, for any configuration $\omega \in \Lambda$ is given by [27, 28]

$$\mathcal{H}(\omega) = \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} J_{ij} (s_i)_\omega (s_j)_\omega \tag{5.2}$$

where J_{ij} is taken from equation (5.1) which incorporates nearest neighbour interactions.

The probability of configuration $\omega \in \Lambda$ is given by the Boltzmann (Gibbs) distribution

$$P(\omega) = \frac{\exp(-\beta \mathcal{H}(\omega))}{\sum_{\omega \in \Lambda} \exp(-\beta \mathcal{H}(\omega))} \tag{5.3}$$

where $\beta = \frac{1}{K_B T}$ such that K_B is the Boltzmann constant and T is temperature. β is very important since this is how the temperature effects the probability. The strength of the Boltzmann distribution lies in the fact that for small values of β (high temperature) the distribution tends to be uniform and for large values of β (low temperature) the probabilities of lowest energy state is accentuated [28, 77]. Therefore, the effective interaction strength increases or decreases depending on temperature T (through β) that in turn effects the probability of the sites being in certain configurations.

5.2 Simulating the Ising model

We shall first give a brief outline of the Metropolis Monte Carlo (MMC) before discussing how we simulated the Ising model using this algorithm. Subsequently we discuss the estimations of probabilities and transition probabilities using temporal averages.

5.2.1 Metropolis Monte Carlo (MMC) algorithm

The definition of Ising model contains no information on its dynamics. However, what one does know is the fraction of the system in a particular configuration (microstate) which is given by the Boltzmann distribution in equation (5.3). The algorithm proposed by Metropolis in 1953 was designed to sample the Boltzmann distribution by artificially imposing

dynamics on the Ising model [62, 27, 77]. This is done by controlling the transition probabilities from one configuration $\omega \in \Lambda$ to another $\omega' \in \Lambda$ such that

$$\gamma_B = P(\omega \to \omega') = \begin{cases} \exp[-\beta(\mathcal{H}(\omega) - \mathcal{H}(\omega'))], & \text{if } \mathcal{H}(\omega') > \mathcal{H}(\omega) \\ 1, & \text{if } \mathcal{H}(\omega') \leq \mathcal{H}(\omega). \end{cases} \tag{5.4}$$

The Hamiltonian $\mathcal{H}$ is taken as in equation (5.2) such that it represents the state of energy at the particular configuration. The transition probability γ_B is favoring lower energy configurations. The Metropolis algorithm can be summarized in these few steps [62, 27]. First prepare the system in an arbitrary configuration ω and calculate $\mathcal{H}(\omega)$. Afterwards choose a random site j so that one can calculate $\mathcal{H}(\omega')$, where ω' is the configuration that will be obtained from ω by changing the state of j such that $(s_j)_\omega = -(s_j)_{\omega'}$ and $(s_i)_\omega = (s_i)_{\omega'}$ for any $i \in \mathcal{N}, i \neq j$. The change on site j is accepted with probability γ_B given by equation (5.4). The process of choosing a new site to flip (upwards or downwards) and comparing the resulting Hamiltonian is then repeated.

Monte Carlo simulations are Markov processes. Based on the Markov chain Monte Carlo procedure, the simulation is primely interested is the invariant distribution of the Markov chain (another name for Markov processes) and not the chain itself [77]. The Metropolis algorithm is Markovian in a sense that the transition probability only depends on the current configuration ω to decide the next configuration. Therefore the product of the Metropolis algorithm is a Markov chain. The most important point is that, Markov chains have invariant distribution. The validity of the Metropolis algorithm depends on the attainability of this invariant distribution (sometimes also known as stationary distribution or the steady state). In the Metropolis algorithm, the intended invariant distribution [77] is the Bolzmann distribution in equation (5.3) which is incorporated into the algorithm as the ratio

$$\frac{P(\omega)}{P(\omega')} = \frac{\exp(-\beta\mathcal{H}(\omega))}{\exp(-\beta\mathcal{H}(\omega'))} = \exp[-\beta(\mathcal{H}(\omega) - \mathcal{H}(\omega'))]$$

in equation (5.4). However the choice of transition probability γ_B in equation (5.4) is not unique. There are other existing choices that may also satisfy the Boltzmann distribution [27].

The implementation of the MMC algorithm in this thesis is outlined as follows. Recall that L is the length of the lattice so that $N = L^2$ is the number of sites on the lattice. A site is chosen at random (in our case using MATLAB's random number generator) to be considered for flipping (change) with probability γ_B. The event of considering the change and afterwards the actual change (if accepted) of the configuration, shall henceforth be referred to as flipping consideration. A sample (or sweep) is taken after each N flipping considerations. The logic being that, since sites to be considered are chosen randomly one at a time, we can assume that after N flips, each site has been selected for consideration once. These samples are the values that we shall refer to when talking about time steps of the resulting Markov chain.

The interaction strength is set to be $J = 1$ and the Boltzmann constant is fixed as $K_B = 1$ for all the simulations. For illustration purposes, $L = 10$ is usually utilized unless stated otherwise. We let the system run up to 2000 samples before sampling at every $N = L^2$ time steps. This is done for more than 100 temperature values of T ranging from 0 to 5.

5.2.2 Temporal average

On the Ising lattice, when one wants to talk about expectations, it must be under the Boltzmann distribution. Therefore for any site X on the lattice, the expectation of $s_X \in \{-1, 1\}$ is given by

$$E_{P(\omega)}[s_X] = \sum_{\omega \in \Lambda} (s_X)_\omega P(\omega) = (1) \sum_{\omega_1 \in \Lambda} P(\omega_1) + (-1) \sum_{\omega_2 \in \Lambda} P(\omega_2), \tag{5.5}$$

where $P(\omega)$ is the probability of the existence of configuration (microstate) w given by the Boltzmann distribution in equation (5.3). ω_1 are the configurations where $s_X = 1$ and ω_2 are the configurations where $s_X = -1$. The average considered here is known in statistical mechanics as the ensemble average [27]. Under certain conditions given by the ergodic theorem, the ensemble average is equal to the temporal average [77, 27]. The temporal average is where the probability of a variable is obtained by averaging over the frequencies of different states over time.

When one generates the values of the Ising model using MMC algorithm, a Markov chain (process) is formed for every site on the lattice. Therefore rather than focusing the state of a site at each configuration $(s_X)_\omega$ (where the configuration has to be taken into account), one focuses on the state of a site at each time step (as obtained from the samples) of the Markov chain $(s_X)_n$ where n is the time step of the Markov chain. Let $\mathcal{T}$ be the length of the Markov chains. To get the temporal average from the Markov chains, one simply counts the frequency of a certain state and then divide it with the length of the Markov chain such that for any $\alpha \in \{-1, 1\}$,

$$P(s_X = \alpha) = \frac{\sum_{n=1}^{\mathcal{T}} \delta\{(s_X)_n = \alpha\}}{|\mathcal{T}|} = p_{s_X}(\alpha) \tag{5.6}$$

where δ is the function defined as:

$$\delta\{.\} = \begin{cases} 1, & \text{if the statement in } \{\ \} \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

For joint distributions we count the joint frequencies, so that for any $\alpha, \beta \in \{-1, 1\}$,

$$P(s_X = \alpha, s_Y = \beta) = \frac{\sum_{n=1}^{\mathcal{T}} \delta\{(s_X)_n = \alpha \text{ and } (s_Y)_n = \beta)\}}{|\mathcal{T}|} = p_{s_X s_Y}(\alpha, \beta). \tag{5.7}$$

The Markov chains generated by the Metropolis Monte Carlo algorithm are known to be able to achieve invariant distribution as well as ergodicity [77]. Consequently, the expectation of s_X can now be written as

$$E_{P(\omega)}[s_X] = \sum_{\omega \in \Lambda} (s_X)_\omega P(\omega) = \sum_{\alpha = \pm 1} (s_X) p_{s_X}(\alpha) \tag{5.8}$$

and the joint expectation would be

$$E_{P(\omega)}[s_X s_Y] = \sum_{\omega \in \Lambda} (s_X)_\omega (s_Y)_\omega P(\omega) = \sum_{\alpha = \pm 1} \sum_{\beta = \pm 1} s_X s_Y p_{s_X s_Y}(\alpha, \beta). \tag{5.9}$$

In relation to equation (1.2), the covariance on the Ising model could be given as

$$\begin{aligned}\Gamma(X,Y) = \Gamma(s_X, s_Y) &= E_{P(\omega)}[s_X s_Y] - E_{P(\omega)}[s_X]E_{P(\omega)}[s_Y] \\ &= \sum_{\alpha=\pm1}\sum_{\beta=\pm1} s_X s_Y p_{s_X s_Y}(\alpha,\beta) - \sum_{\alpha=\pm1}(s_X)p_{s_X}(\alpha)\sum_{\beta=\pm1}(s_Y)p_{s_Y}(\beta) \\ &= \sum_{\alpha=\pm1}\sum_{\beta=\pm1} s_X s_Y \left[p_{s_X s_Y}(\alpha,\beta) - p_{s_X}(\alpha)p_{s_Y}(\beta)\right], \end{aligned} \tag{5.10}$$

where $X, Y \in N$ are two sites on the lattice and the temporal average is applied. All the numerical probabilities obtained for Ising model in this thesis will have been obtained using the temporal average method on MMC simulations where the resulting Markov chains will be of length $\mathcal{T} = 100000$ unless stated otherwise.

5.2.3 Estimating transition probabilities

To get transition probabilities, again we utilize the fact that under the Metropolis Monte Carlo simulations each site is considered as a Markov chain. One way of doing this is to simply count and get the fraction of the occurance of these transition. For example to get the transition probability of s_X from state β to state α at time lag 1, we have

$$P((s_X)_n = \alpha | (s_X)_{n-1} = \beta) = \frac{\sum_{n=2}^{\mathcal{T}} \delta\{(s_X)_n = \alpha \text{ given } (s_X)_{n-1} = \beta)\}}{|\mathcal{T}-1|}.$$

where δ is again the function defined in equation (5.6). In other words, we use temporal average to obtain the transition probabilities. Generally for any time lag τ one can calculate the transition probability such that for any $\alpha, \beta \in \{-1, 1\}$

$$P((s_X)_n = \alpha | (s_X)_{n-\tau} = \beta) = \frac{\sum_{n=1+\tau}^{\mathcal{T}} \delta\{(s_X)_n = \alpha \text{ given } (s_X)_{n-\tau} = \beta)\}}{|\mathcal{T}-\tau|}.$$

However if this is done, the marginal probability in equation (5.6) will also need to be altered so that the probabilities will tally. The marginal probability for any τ is

$$P(s_X = \alpha) = \frac{\sum_{n=1+\tau}^{\mathcal{T}} \delta\{(s_X)_n = \alpha\}}{|\mathcal{T}-\tau|} = p_{s_X}(\alpha). \tag{5.11}$$

Consequently, when using this method one needs to recalculate the marginal probabilities for each different value of time lag τ. In order to avoid the need to recalculate at every time lag, one can utilize a different method.

Alternatively, consider X^τ to be a different Markov chain where X^τ is the process X that is shifted by τ time steps. In doing this, we do not change the length of the Markov chain but instead we shift the Markov chain in a circular manner so that the marginal probabilities remain the same for any time lag τ. In other words, the first τ time steps of X becomes the last τ time steps of X^τ such that $(s_{X^\tau})_{\mathcal{T}-\tau+n} = (s_X)_n$ for $n = 1, \cdots, \tau$, and $(s_{X^\tau})_{n-\tau} = (s_X)_n$ for $n = \tau + 1, \cdots, \mathcal{T}$. We simply let $Y = X^\tau$ in equation (5.7) so that

$$P(s_X = \alpha, s_{X^\tau} = \beta) = \frac{\sum_{n=1}^{\mathcal{T}} \delta\{(s_X)_n = \alpha \text{ and } (s_{X^\tau})_n = \beta)\}}{|\mathcal{T}|} = p_{s_X s_{X^\tau}}(\alpha, \beta). \quad (5.12)$$

In both these methods, it is important that $\tau << \mathcal{T}$ so that the probability estimation obtained from the simulated Markov chain is as accurate as possible.

5.3 Measures on Ising model

Utilizing the time average method of approximation, we are able to approximate the measures that we have defined in previous chapters on the Ising model and some new ones due to the nature of the Ising model. First we explore the measures or observables that may be of interest due to their relationship with phase transition and critical values on the Ising model. After that, we revisit covariance and various forms of Mutual Information including conditional Mutual Information and Transfer Entropy.

5.3.1 Observables for verification of the critical point

In an infinite two dimensional lattice, the phase transition of the Ising model with $J = 1$ and $K_B = 1$ is known to occur at the critical temperature $T_c = \frac{2}{log(1+\sqrt{2})} \approx 2.269185$ [27]. In a finite system, due to finite size effects, the critical values will not be quite as exact, we will refer the temperature where the transition occurs in the simulation as the crossover temperature T_c. Magnetisation M and susceptibility χ are observables that are normally used to identify T_c on the Ising model.

In order to define M and χ, let $m(n) = \sum_{i=1}^{N}(s_i)_n$ be the sum of spins on a lattice of size N at time steps $n = 1, \cdots, \mathcal{T}$. The magnetisation M is defined as

$$M = \frac{1}{N}E[m(n)] = \frac{1}{N}\left[\frac{1}{\mathcal{T}}\sum_{n=1}^{\mathcal{T}} m(n)\right] = \frac{1}{N\mathcal{T}}\sum_{n=1}^{\mathcal{T}}\sum_{i=1}^{N}(s_i)_n. \tag{5.13}$$

We utilized $E[.]$ in terms of temporal average by averaging over all the time steps n. Subsequently, with $K_B = 1$ the susceptibility per spin [27] can be written as

$$\begin{aligned}\chi &= \frac{1}{TN}\left(E[m(n)^2] - E[m(n)]^2\right) \\ &= \frac{1}{TN}\left(\frac{1}{\mathcal{T}}\sum_{n=1}^{\mathcal{T}}\left[\sum_{i=1}^{N}(s_i)_n\right]^2 - \left[\frac{1}{\mathcal{T}}\sum_{n=1}^{\mathcal{T}}\sum_{i=1}^{N}(s_i)_n\right]^2\right)\end{aligned} \tag{5.14}$$

where T is the temperature. Using MMC algorithm for temperatures $T = 0, \cdots, 5$ (taking $\chi = 0$ at $T = 0$) and chain length (number of samples for each site) $\mathcal{T} = 100000$ we get M in Figure (5.1) and χ in Figure (5.2). The M values in Figure (5.1) was initialized with values 1 (all spins-up) therefore the initial magnetisation is 1. If the initialization was with spins-down values (-1), then the initial magnetisation would have been -1.

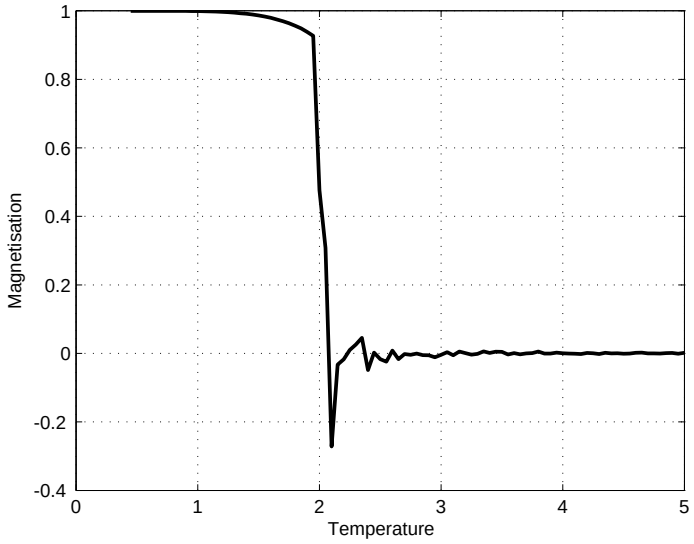

Figure 5.1: Magnetisation M using equation (5.13) stabilizes to 0 at T_c

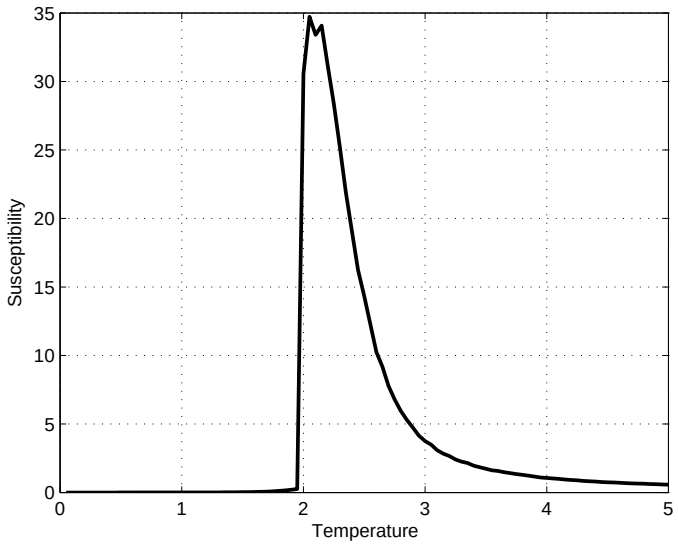

Figure 5.2: Susceptibility χ of equation (5.14) peaks at T_c

The Mutual Information between sites X and Y on the lattice, in relation with equation (2.12), (5.6) as well as (5.7), can be written as

$$\begin{aligned} I(X,Y) = I(s_X, s_Y) &= E\left[log\frac{P(s_X,s_Y)}{P(s_X)P(s_Y)}\right] \\ &= \sum_{\alpha=\pm1}\sum_{\beta=\pm1} p_{s_X s_Y}(\alpha,\beta)\, log\frac{p_{s_X s_Y}(\alpha,\beta)}{p_{s_X}(\alpha)p_{s_Y}(\beta)}. \end{aligned} \tag{5.15}$$

Consequently conditional Mutual Information involving another site Z can be written as

$$\begin{aligned} I(X,Y|Z) = I(s_X, s_Y|s_Z) &= E\left[log\frac{P(s_X,s_Y|s_Z)}{P(s_X|s_Z)P(s_Y|s_Z)}\right] \\ &= \sum_{\alpha=\pm1}\sum_{\beta=\pm1}\sum_{\gamma=\pm1} p_{s_X s_Y s_Z}(\alpha,\beta,\gamma)\, log\frac{p_{s_X s_Y|s_Z}(\alpha,\beta|\gamma)}{p_{s_X|s_Z}(\alpha|\gamma)p_{s_Y|s_Z}(\beta|\gamma)}. \end{aligned} \tag{5.16}$$

To see the effect of Mutual Information over different temperatures we choose three sites A, B and G representing coordinates $[1,1]$, $[2,2]$ and $[3,3]$ on the lattice. In a lattice with $L = 10$ with $N = 100$ sites, we have that $A = 1, B = 12$ and $G = 23$ so that $A, B, G \in \mathcal{N}$. Matsuda [73] concluded that the Mutual Information and covariance shows singular behaviour near critical point and this is what we observe in Figure (5.3).

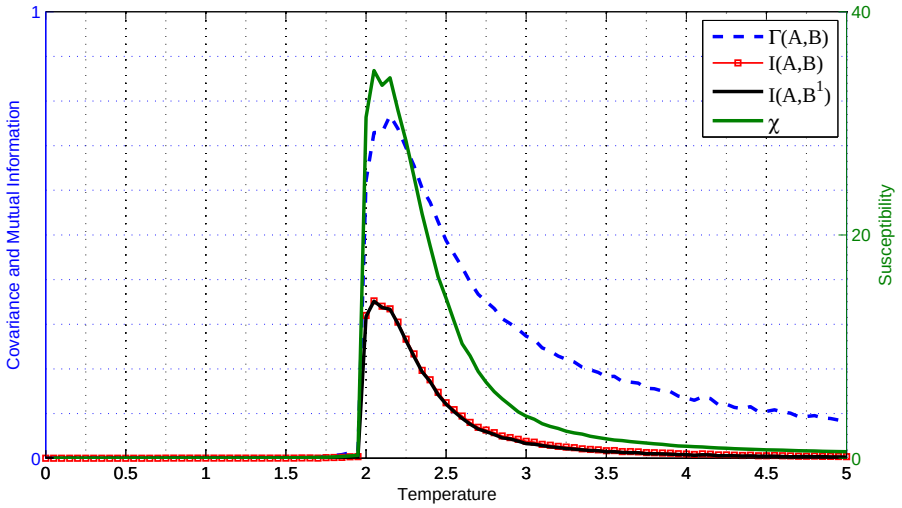

Figure 5.3: Values of covariance $\Gamma(A, B)$ using equation (5.10), Mutual Information $I(A, B)$ using equation (5.15), time delayed Mutual Information $I(A, B^1)$ using equation (5.17) and susceptibility χ using equation (5.14) on the simulated values of the Ising model across temperature T

5.3.2 Measuring values across time lags

In Figure (5.3) we have plotted time delayed Mutual Information $I(A, B^1)$ which is almost indistinguishable from Mutual Information $I(A, B)$. Utilizing the probability estimation in equation (5.6) and (5.12), the time delayed Mutual Information between any site X and Y on the lattice can be defined as

$$\begin{aligned} I(X, Y^\tau) = I(s_X, s_{Y^\tau}) &= E\left[log\frac{P(s_X, s_{Y^\tau})}{P(s_X)P(s_{Y^\tau})}\right] \\ &= \sum_{\alpha=\pm1}\sum_{\beta=\pm1} p_{s_X s_{Y^\tau}}(\alpha, \beta)\, log\frac{p_{s_X s_{Y^\tau}}(\alpha, \beta)}{p_{s_X}(\alpha)p_{s_{Y^\tau}}(\beta)} \end{aligned} \tag{5.17}$$

where Y^τ is the variable Y shifted by τ time steps. Previously we have also defined auto Mutual Information or the Mutual Information over time as $I(X, X^\tau)$, this is a special case of time delayed Mutual Information on itself. From our investigations on the simulated data, neither time delayed nor auto Mutual Information values will be very different from Mutual Information values on the model.

The Transfer Entropy on the Ising model can be calculated using the equation (4.5)

$$T_{YX}^{(\tau)} = T_{s_Y s_X}^{(\tau)} = E\left[log\frac{P(s_X|s_{X^{-1}}, s_{Y^{-\tau}})}{P(s_X|s_{X^{-1}})}\right]$$
$$= \sum_{\alpha=\pm1}\sum_{\beta=\pm1}\sum_{\gamma=\pm1} p_{s_X s_{X^{-1}} s_{Y^{-\tau}}}(\alpha,\beta,\gamma)\, log\frac{p_{s_X|s_{X^{-1}}s_{Y^{-\tau}}}(\alpha|\beta,\gamma)}{p_{s_X|s_{X^{-1}}}(\alpha|\beta)}. \quad (5.18)$$

In Figure (5.4) we plot the values $T_{BA}^{(1)}$ and $T_{AB}^{(1)}$ alongside susceptibility χ. Despite the

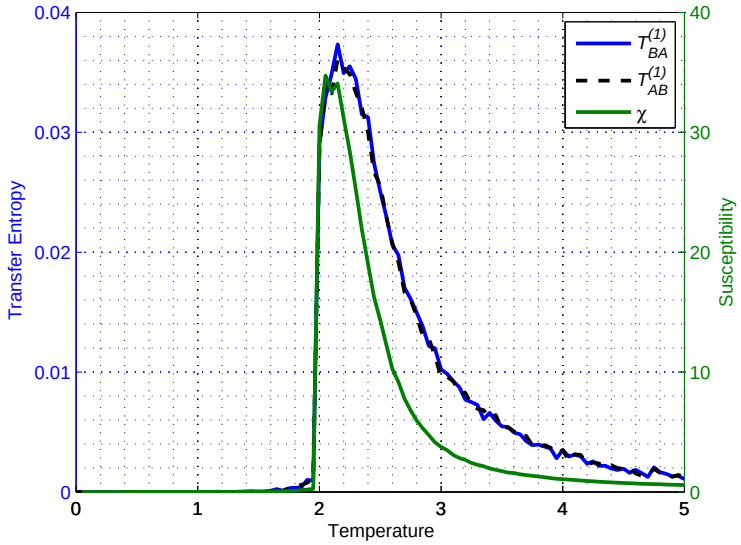

Figure 5.4: Values of $T_{BA}^{(1)}$ and $T_{AB}^{(1)}$ both using equation (5.18) and susceptibility χ using equation (5.14), across temperature T.

smaller values of Transfer Entropy it still does peak near T_c. One can see that there is no clear difference between $I(A,B)$ and $I(A,B^1)$ in Figure (5.3) nor between $T_{BA}^{(1)}$ and $T_{AB}^{(1)}$ in Figure (5.4), thus no direction of 'causality' can be established between A and B. This is true for any τ used in equations (5.17) and (5.18) between any two site on the lattice. What we observed was the effect of the distance between the sites.

5.3.3 The influence of distance

In this subsection we shall use reduced temperature $\frac{T-T_c}{T_c}$ for visualization. From M in Figure (5.1) and χ in Figure (5.2), the crossover value is estimated to be $T_c = 2.15$. The

effect of distance is observed in the values of Mutual Information and covariance as seen in Figures (5.5) and (5.6).

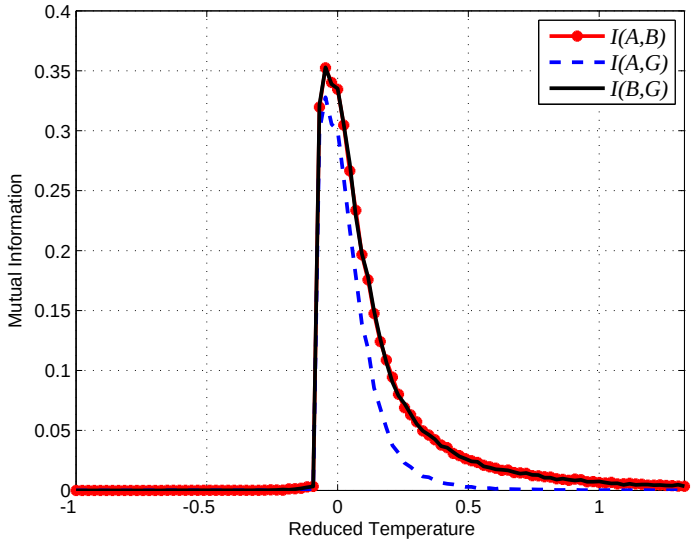

Figure 5.5: Mutual Information $I(A,B)$, $I(A,G)$ and $I(B,G)$ of equation (5.15) versus reduced temperature $\frac{T-T_c}{T_c}$. $I(A,G) < I(A,B) \approx I(B,G)$ due to distance.

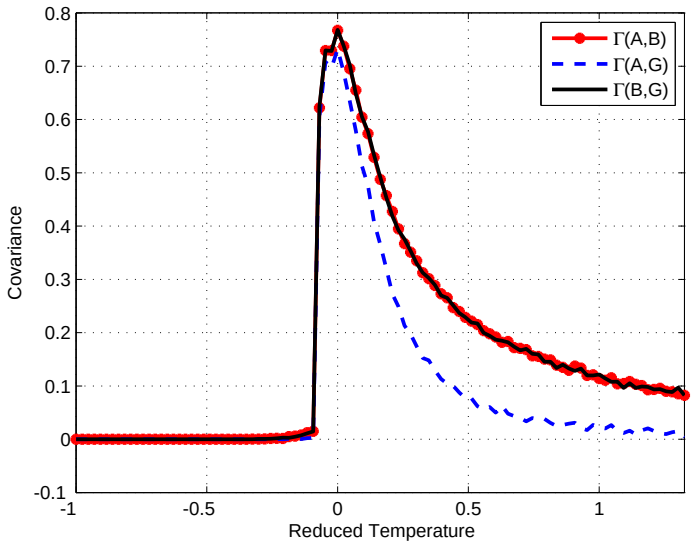

Figure 5.6: Covariance $\Gamma(A,B)$, $\Gamma(A,G)$ and $\Gamma(B,G)$ of equation (5.18) versus reduced temperature $\frac{T-T_c}{T_c}$. $\Gamma(A,G) < \Gamma(A,B) \approx \Gamma(B,G)$ due to distance.

The fact that A, B and G represents coordinates $[1, 1]$, $[2, 2]$ and $[3, 3]$ means that the distance from A to B and B to G is equal but half the distance of A to G. Due to strictly nearest neighbour interactions as well as the translational and rotational invariance nature of the Ising model, two sites on the lattice with the same distance between them will have the same marginal and joint probability. This is evident from how $I(A, B) \approx I(B, G)$ in Figure (5.5) and $\Gamma(A, B) \approx \Gamma(B, G)$ in Figure (5.6). The values of $I(A, G)$ and $\Gamma(A, G)$ are smaller in the respective figures due to the larger distances between the sites.

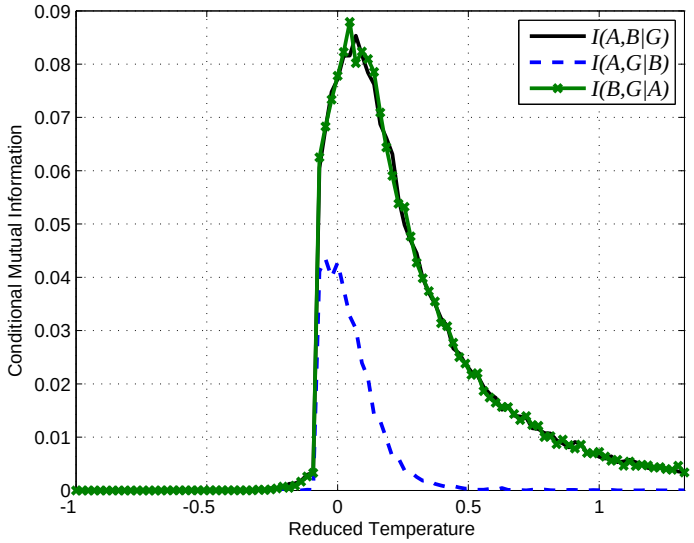

Figure 5.7: Conditional Mutual Information $I(A, B|G)$, $I(A, G|B)$ and $I(B, G|A)$ of equation (5.16) versus reduced temperature $\frac{T-T_c}{T_c}$. $I(A, G) < I(A, B) \approx I(B, G)$ due to distance.

B is situated between A and G on the lattice. Although $[2, 2]$ is not the nearest neighbour of either $[1, 1]$ or $[3, 3]$, interactions between to the two sites will logically pass through $[2, 2]$. In a way, this makes the interaction between A and G dependent on B. When the conditional Mutual Information values between the three sites A, B and G are plotted in Figure (5.7), among the three values, $I(A, G|B)$ is the smallest. This is not only due to the fact that A and G is further away from each other, but also because B which is situated between the other two sites is conditioned out in $I(A, G|B)$. Therefore one can say that the conditional Mutual Information uncovers indirect dependence in a sense that the interaction between A and G depends on B. We have discussed a possible relationship between

conditional Mutual Information and indirect 'causality' in subsection (4.3.2).

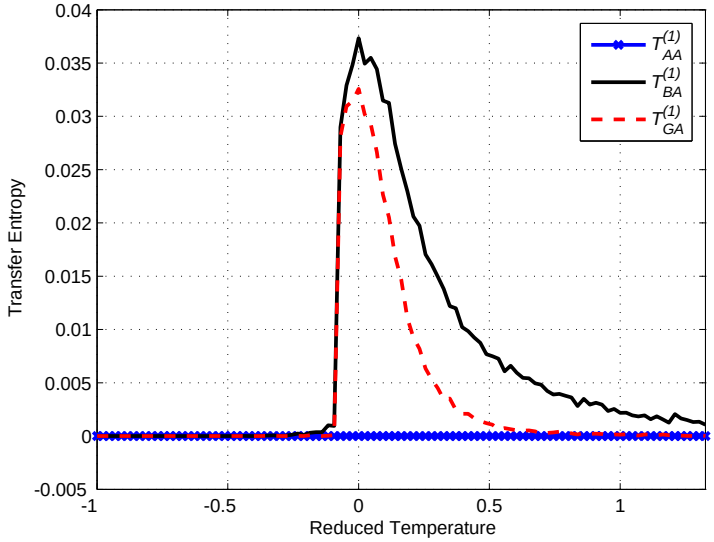

Figure 5.8: Transfer Entropy $T^{(1)}_{AA}$, $T^{(1)}_{BA}$ and $T^{(1)}_{GA}$ of equation (5.18) versus reduced temperature $\frac{T-T_c}{T_c}$. $T^{(1)}_{GA} < T^{(1)}_{BA}$ due to distance.

The Transfer Entropy and the conditional Mutual Information is related by $T^{(\tau)}_{YX} = I(X, Y^{-\tau}|X^{-1})$ as in equation (4.6), therefore the distance and conditioning will also effect the values of Transfer Entropy. In Figure (5.8), one can see that $T^{(1)}_{AA} = I(A, A^{-1}|A^{-1}) = 0$ and $T^{(1)}_{BA} = I(A, B^{-1}|A^{-1}) > T^{(1)}_{GA} = I(A, G^{-1}|A^{-1})$. Therefore it can be said that distance is the only factor effecting Transfer Entropy values in this figure and from Figure (5.4) there seems to be no particular causal direction either. We suspect that this is due to the symmetric nature of the Ising model that distributes influences equally in all direction.

5.3.4 Measures on $L = 25$

Up to this point we have utilized lattice length of $L = 10$. In this subsection we contrast $L = 10$ to $L = 25$ with Markov chain lengths of $\mathcal{T} = 100000$. In Figure (5.9) we observe that the value of χ increases as L increases since $\chi \to \infty$ as $L \to \infty$. The crossover temperature of $L = 25$ is $T_c \approx 2.2$ which is closer to the real T_c. In a lattice of length $L = 25$ with $N = 625$ sites, sites A, B and G representing coordinates $[1, 1]$, $[2, 2]$ and $[3, 3]$ will have values of $A = 1, B = 27$ and $G = 53$ such that $A, B, G \in \mathcal{N}$.

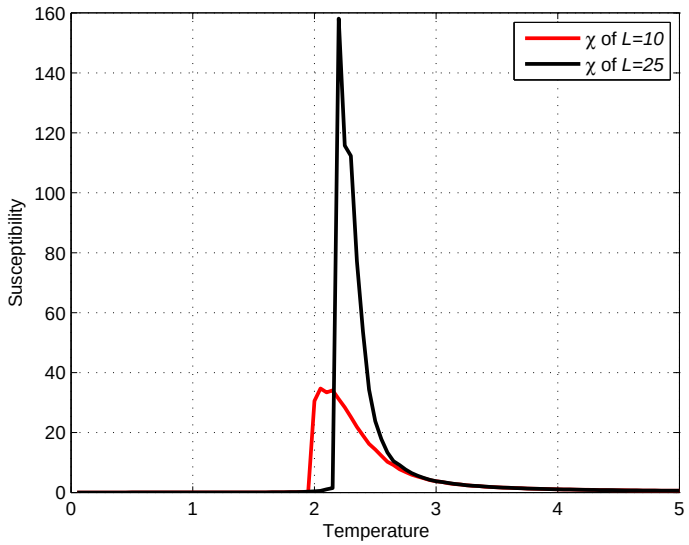

Figure 5.9: Susceptibility χ of equation (5.14) for lattices of lengths $L = 10, 25$.

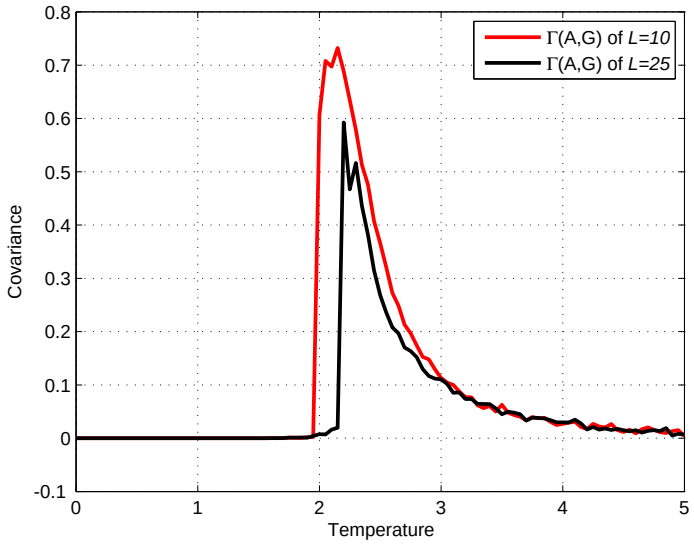

Figure 5.10: Covariance $\Gamma(A, G)$ in equation (5.10) for lattices of lengths $L = 10, 25$.

Figures (5.10), (5.11) and (5.12) depicts the behaviour of covariance, Mutual Information and Transfer Entropy values across temperatures $T = 0 \cdots 5$ between site A at coordinate $[1, 1]$ and site G at coordinate $[3, 3]$ in all three different lattices. We suspect that the explanation to what we observe in these graph is that for $L = 25$ where the lattice

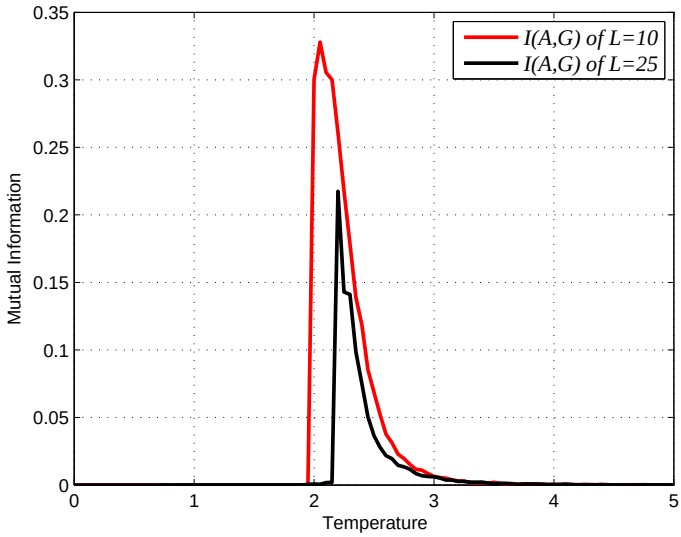

Figure 5.11: Mutual Information $I(A, G)$ using equation (5.15) on simulated data of lattices with lengths $L = 10, 25$.

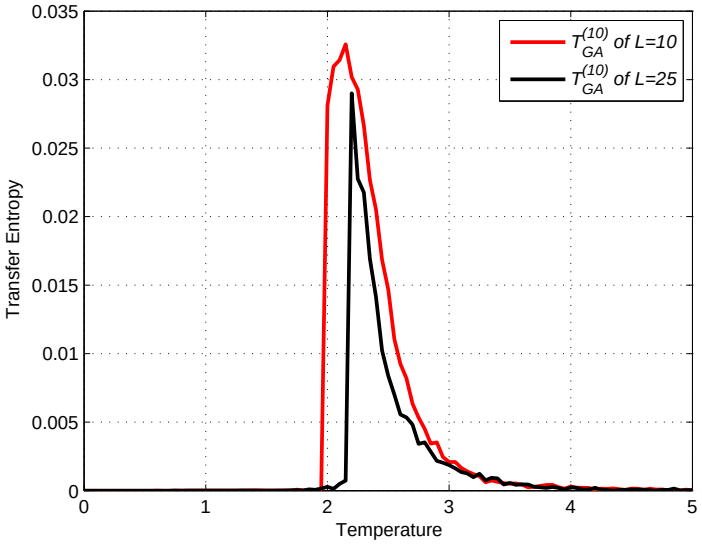

Figure 5.12: $T_{GA}^{(10)}$ using equation (5.18) on simulated lattices with lengths $L = 10, 25$.

is much bigger, the influence of the two sites on each other is much weaker than in the $L = 10$ lattice especially with periodic boundary boundary conditions, hence the sharper and more precise detection of T_c in the $L = 25$ lattice. Nevertheless, the fact that the measures attained maximum values near T_c is consistent to our observation on $L = 10$.

5.4 Binary sequences

A binary sequence or a Bernoulli process is a process that only allows only two possible states. The fact that Ising model is a binary sequence simplifies a lot, so much so that even Ising himself wrongly came to the conclusion that his model was of no use. Its contribution to the world statistical physics is undeniable, however when we look closer we find that in binary sequences covariance Γ and Mutual Information I are in fact interchangeable in terms of independence. A relationship between the two measures on Ising model will be used to explain Figure (5.3). We will see that given a few assumptions, on a binary sequence, the linear independence of the covariance will be enough to indicate general independence.

5.4.1 Independence of binary sequence 0 and 1

We leave the Ising model for a while in the quest to explain things in the more general setting. There are some special cases where uncorrelated-ness does imply general independence. One of these cases is when the variables have only two possible values i.e the binary sequence. The binary sequence is usually represented by 0 and 1. Let $p_X(x)$ and $p_Y(y)$ be the marginal probabilities and $p_{XY}(x,y)$ be the joint probability for variables X and Y where $x, y \in \{0,1\}$. The covariance in equation (1.2) becomes

$$\Gamma = \Gamma(X,Y) = p_{XY}(1,1) - p_X(1)p_Y(1). \tag{5.19}$$

If X and Y are uncorrelated ($\Gamma = 0$) then $p_{XY}(1,1) = p_X(1)p_Y(1)$ and using this in relation to the property of joint probabilities $\sum_y p_{XY}(x,y) = p_X(x)$ gives us

$$\begin{aligned} p_X(1) &= p_{XY}(1,1) + p_{XY}(1,0) \\ &= p_X(1)p_Y(1) + p_{XY}(1,0) \\ p_X(1)[1 - p_Y(1)] &= p_{XY}(1,0) \\ p_X(1)p_Y(0) &= p_{XY}(1,0). \end{aligned} \tag{5.20}$$

Proceeding in a similar manner, it is possible to get $p_X(0)p_Y(1) = p_{XY}(0,1)$ as well as $p_X(0)p_Y(0) = p_{XY}(0,0)$, so that $p_{XY}(x,y) = p_X(x)p_Y(y)$ for all $x \in \{0,1\}$ and $y \in \{0,1\}$, making X and Y independent of each other based on definition of general independence. Therefore linear independence implies general independence when $x, y \in \{0,1\}$. Consequently, $\Gamma(X,Y) = 0 \Rightarrow I(X,Y) = 0$ in this case.

We have briefly discussed a similar example in subsection (2.3.2) but the independence was not explicitly highlighted. In subsection (2.3.2), a formula that links Γ to I have been obtained by writing the probabilities in terms of $\Gamma = \Gamma(X,Y)$. This was proposed by [64], which imposes symmetric condition $p_{XY}(0,1) = p_{XY}(1,0)$ so that using the joint probabilities we get $p_X(0) = p_Y(0)$, $p_X(1) = p_Y(1)$ and $p_X(1) - p_{XY}(1,1) = p_X(0) - p_{XY}(0,0)$. We highlight here that even without the symmetric condition, the independence between X and Y have already been establish.

Taking into account $p_X(1) + p_X(0) = 1$ (the normalizing condition) and substituting values in equation (5.19) yields

$$p_{XY}(1,1) = \Gamma + p_X(1)^2, \tag{5.21a}$$

$$p_{XY}(0,0) = \Gamma + p_X(0)^2, \tag{5.21b}$$

$$p_{XY}(0,1) = p_{XY}(1,0) = -\Gamma + p_X(0)p_X(1). \tag{5.21c}$$

The probabilities can be used to obtain the Mutual Information formula [64] by substituting the probabilities into equation (2.12) such that

$$\begin{aligned} I(X,Y) = &\sum_{x=\pm 1}\sum_{y=\pm 1} p_{XY}(x,y)\, log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \\ =&\Gamma\ log\frac{\left(1+\frac{\Gamma}{p_X(1)^2}\right)\left(1+\frac{\Gamma}{p_X(0)^2}\right)}{\left(1-\frac{\Gamma}{p_X(0)p_X(1)}\right)^2} + p_X(1)^2 log\left(1+\frac{\Gamma}{p_X(1)^2}\right) \\ &+ p_X(0)^2 log\left(1+\frac{\Gamma}{p_X(0)^2}\right) + 2p_X(0)p_X(1) log\left(1-\frac{\Gamma}{p_X(0)p_X(1)}\right). \end{aligned} \tag{5.22}$$

When the terms $\frac{\Gamma}{p_X(x)p_Y(y)}$ are small, one common method of approximation is to use the

second order Taylor approximation in obtaining

$$\begin{aligned} log\left(1+\frac{\Gamma}{p_X(1)^2}\right) &\approx \frac{\Gamma}{p_X(1)^2}\left[1-\frac{\Gamma}{2p_X(1)^2}\right] \\ log\, g\left(1+\frac{\Gamma}{p_X(0)^2}\right) &\approx \frac{\Gamma}{p_X(0)^2}\left[1-\frac{\Gamma}{2p_X(0)^2}\right] \\ log\left(1-\frac{\Gamma}{p_X(0)p_X(1)}\right) &\approx -\frac{\Gamma}{p_X(0)p_X(1)}\left[1+\frac{\Gamma}{2p_X(0)p_X(1)}\right]. \end{aligned} \tag{5.23}$$

Taking only the second order terms of Γ (the first order cancels out), we have

$$\begin{aligned} I \approx & \frac{\Gamma^2}{2}\left(\frac{1}{p_X(1)^2}+\frac{1}{p_X(0)^2}+\frac{2}{p_X(0)p_X(1)}\right) \\ = & \frac{\Gamma^2}{2}\left(\frac{p_X(1)^2+p_X(0)^2+2p_X(0)p_X(1)}{\left[p_X(0)p_X(1)\right]^2}\right) \\ = & \frac{\Gamma^2}{2}\left(\frac{p_X(1)+p_X(0)}{p_X(0)p_X(1)}\right)^2 = \frac{1}{2}\left(\frac{\Gamma}{p_X(0)p_X(1)}\right)^2. \end{aligned} \tag{5.24}$$

What we can observe from this equation is that I decays to zero at a faster rate than the corresponding Γ and more importantly we can clearly see the fact that $I = 0 \Leftrightarrow \Gamma = 0$ for this particular approximation. It has to be said that this is not true as soon as one goes into ternary sequences. Examples of $\Gamma = 0 \not\Rightarrow I = 0$ beyond binary have been discussed in subsection (2.3.2) and [64].

5.4.2 Covariance and Mutual Information for general binary sequence

We show how the works of [64] extends to general binary sequence with states α and β. Now, define $\tilde{X}$ and $\tilde{Y}$ just like X and Y except that $\tilde{x}, \tilde{y} \in \{\alpha, \beta\}$ as opposed to $x, y \in \{0, 1\}$. Let state α correspond to 0 and state β correspond to 1 so that same distributions are maintained i.e. $p_{\tilde{X}}(\alpha) = p_X(0)$, $p_{\tilde{X}\tilde{Y}}(\alpha, \beta) = p_{XY}(0, 1)$ on so on (also imposing the

symmetry constraint). Using the fact that $\frac{\alpha-\alpha}{\beta-\alpha}=0$ and $\frac{\beta-\alpha}{\beta-\alpha}=1$ we have $\forall\alpha,\beta,\alpha\neq\beta$,

$$\begin{aligned}\Gamma\left(\frac{\tilde{X}-\alpha}{\beta-\alpha},\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)=&E\left(\frac{\tilde{X}-\alpha}{\beta-\alpha}\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)-E\left(\frac{\tilde{X}-\alpha}{\beta-\alpha}\right)E\left(\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)\\ =&\sum_{\tilde{x}}\sum_{\tilde{y}}\left(\frac{\tilde{x}-\alpha}{\beta-\alpha}\frac{\tilde{y}-\alpha}{\beta-\alpha}\right)p_{\tilde{X}\tilde{Y}}(\tilde{x},\tilde{y})\\ &-\sum_{\tilde{x}}\left(\frac{\tilde{x}-\alpha}{\beta-\alpha}\right)p_{\tilde{X}}(\tilde{x})\sum_{\tilde{y}}\left(\frac{\tilde{y}-\alpha}{\beta-\alpha}\right)p_{\tilde{Y}}(\tilde{y})\\ =&\sum_{x}\sum_{y}(xy)\,p_{XY}(x,y)-\sum_{x}(x)\,p_X(x)\sum_{y}(y)\,p_Y(y)\\ =&\Gamma(X,Y)=p_{XY}(1,1)-p_X(1)p_Y(1).\end{aligned}\tag{5.25}$$

Therefore, the relationship between $\Gamma(X,Y)$ and $\Gamma(\tilde{X},\tilde{Y})$ is such that

$$\begin{aligned}\Gamma(X,Y)&=\Gamma\left(\frac{\tilde{X}-\alpha}{\beta-\alpha},\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)\\ &=E\left(\frac{\tilde{X}-\alpha}{\beta-\alpha}\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)-E\left(\frac{\tilde{X}-\alpha}{\beta-\alpha}\right)E\left(\frac{\tilde{Y}-\alpha}{\beta-\alpha}\right)\\ &=\frac{E(\tilde{X}\tilde{Y}-\alpha[\tilde{X}+\tilde{Y}]+\alpha^2)-(E(\tilde{X})-\alpha)(E(\tilde{Y})-\alpha)}{(\beta-\alpha)^2}\\ &=\frac{E(\tilde{X}\tilde{Y})-E(\tilde{X})E(\tilde{Y})}{(\beta-\alpha)^2}=\frac{\Gamma(\tilde{X},\tilde{Y})}{(\beta-\alpha)^2}.\end{aligned}\tag{5.26}$$

Since $(\beta-\alpha)^2$ is constant, then $\Gamma(\tilde{X},\tilde{Y})=(\beta-\alpha)^2\,\Gamma(X,Y)$ implies that $\Gamma(\tilde{X},\tilde{Y})$ is proportional to $\Gamma(X,Y)$. Whenever $\Gamma(\tilde{X},\tilde{Y})=0$ we have that $\Gamma(X,Y)=\frac{\Gamma(\tilde{X},\tilde{Y})}{(\beta-\alpha)^2}=0$ since and $\beta\neq\alpha$ (otherwise the variables are just constants), therefore we have independence. In short, for any binary sequence where $\tilde{x},\tilde{y}\in\{\alpha,\beta\}$, linear independence (uncorrelatedness) implies general independence since we have

$$\begin{aligned}\Gamma(\tilde{X},\tilde{Y})=0\Rightarrow\Gamma(X,Y)=0\Rightarrow p_{XY}(x,y)=p_X(x)p_Y(y),x,y\in\{0,1\}\\ \Rightarrow p_{\tilde{X}\tilde{Y}}(\tilde{x},\tilde{y})=p_{\tilde{X}}(\tilde{x})p_{\tilde{Y}}(\tilde{y}),\tilde{x},\tilde{y}\in\{\alpha,\beta\}.\end{aligned}\tag{5.27}$$

Recall that $I(\tilde{X},\tilde{Y}) = I(X,Y)$ since the probabilities are identical and Mutual Information only depends on probabilities [89, 57]. Letting $\tilde{\Gamma} = \tilde{\Gamma}(X,Y)$ and substituting $\Gamma = \frac{\tilde{\Gamma}}{(\beta-\alpha)^2}$ as well as the probabilities in equation (5.22) yields

$$\begin{aligned} I(\tilde{X},\tilde{Y}) =& I(X,Y) \\ =& \frac{\tilde{\Gamma}}{(\beta-\alpha)^2}\ log \frac{\left(1+\frac{\tilde{\Gamma}}{\left[p_{\tilde{X}}(\beta)(\beta-\alpha)\right]^2}\right)\left(1+\frac{\tilde{\Gamma}}{\left[p_{\tilde{X}}(\alpha)(\beta-\alpha)\right]^2}\right)}{\left(1-\frac{\tilde{\Gamma}}{p_{\tilde{X}}(\alpha)p_{\tilde{X}}(\beta)(\beta-\alpha)^2}\right)^2} \\ &+ p_{\tilde{X}}(\beta)^2 log\left(1+\frac{\tilde{\Gamma}}{\left[p_{\tilde{X}}(\beta)(\beta-\alpha)\right]^2}\right) + p_{\tilde{X}}(\alpha)^2 log\left(1+\frac{\tilde{\Gamma}}{\left[p_{\tilde{X}}(\alpha)(\beta-\alpha)\right]^2}\right) \\ &+ 2p_{\tilde{X}}(\alpha)p_{\tilde{X}}(\beta)\, log\left(1-\frac{\tilde{\Gamma}}{p_{\tilde{X}}(\alpha)p_{\tilde{X}}(\beta)(\beta-\alpha)^2}\right). \end{aligned} \tag{5.28}$$

Clearly for any binary sequence with any value of states $\Gamma(\tilde{X},\tilde{Y}) = 0 \Rightarrow\ I(\tilde{X},\tilde{Y}) = 0$ i.e uncorrelated-ness is enough to imply general independence. Approximating as in equation (5.24), we get that

$$\begin{aligned} \tilde{I} \approx& \frac{1}{2}\left(\frac{\Gamma}{p_X(0)p_X(1)}\right)^2 = \frac{1}{2}\left(\frac{\tilde{\Gamma}}{(\beta-\alpha)^2 p_X(0)p_X(1)}\right)^2 \\ =& \frac{1}{2(\beta-\alpha)^4}\left(\frac{\tilde{\Gamma}}{p_X(0)p_X(1)}\right)^2. \end{aligned} \tag{5.29}$$

The relationship between $\tilde{I}$ and $\tilde{\Gamma}$ depends on the difference between the possible binary states $(\beta-\alpha)$. This makes sense since the values of $\tilde{\Gamma}$ will be larger for larger values of α and β.

5.4.3 Ising model as a binary sequence

We now return to the Ising model and its notation. The Ising model is a particular example of the binary sequence where $\alpha = -1$ and $\beta = 1$. However as previously defined, the variables on the Ising model are s_X and s_Y. It is the state of sites X and Y that is considered as the Markov chains. Two sites X and Y on the Ising lattice is said to be independent of each

other when the probabilities $p_{s_X s_Y}(\alpha,\beta) = p_{s_X}(\alpha)p_{s_Y}(\beta), \forall \alpha,\beta \in \{-1,1\}$ which can be obtained by referring to equation (5.27). As special case of the binary sequence, linear independence implies general independence on the Ising model. Therefore, $\Gamma(s_X, s_Y) = 0 \Rightarrow I(s_X, s_Y) = 0$. In conclusion, on the Ising lattice, covariance is sufficient to indicate general independence.

Moreover, given the symmetry condition of the Ising Model which is justified by the fact the Ising model is translational and rotational invariance, one can obtain the Mutual Information in terms of covariance $\Gamma = \Gamma(s_X, s_Y)$ on the Ising model utilizing equation (5.28), such that

$$\begin{aligned} I(s_X, s_Y) =& \frac{\Gamma}{4}\ log \frac{\left(1+\frac{\Gamma}{4\left[p_{s_X}(1)\right]^2}\right)\left(1+\frac{\Gamma}{4\left[p_{s_X}(-1)\right]^2}\right)}{\left(1-\frac{\Gamma}{4p_{s_X}(1)p_{s_X}(-1)}\right)^2} \\ &+ p_{s_X}(1)^2\, log\left(1+\frac{\Gamma}{4\left[p_{s_X}(1)\right]^2}\right) + p_{s_X}(-1)^2\, log\left(1+\frac{\Gamma}{4[p_{s_X}(-1)]^2}\right) \\ &+ 2p_{s_X}(1)p_{s_X}(-1)\, log\left(1-\frac{\Gamma}{4p_{s_X}(1)p_{s_X}(-1)}\right). \end{aligned} \tag{5.30}$$

Clearly $I = 0$ when $\Gamma = 0$. One can also apply the approximation in equation (5.29) to obtain

$$\begin{aligned} I = I(s_X, s_Y) \approx& \frac{1}{2(\beta-\alpha)^4}\left(\frac{\Gamma(s_X,s_Y)}{p_{s_X}(-1)p_{s_X}(1)}\right)^2 = \frac{1}{2(2)^4}\left(\frac{\Gamma(s_X,s_Y)}{p_{s_X}(-1)p_{s_X}(1)}\right)^2 \\ =& \frac{1}{32}\left(\frac{\Gamma(s_X,s_Y)}{p_{s_X}(-1)p_{s_X}(1)}\right)^2. \end{aligned} \tag{5.31}$$

Thus $I \approx \Gamma^2$ as seen in Figure (5.13). The fact the Ising model is translational and rotational invariance enables sites to be grouped by distances such that the probability of two sites with the same distance from each other is taken to be the same as other with the same distance [73] as we have seen in subsection (5.3.3). Plotting equations (5.10) and (5.15) for all $X, Y \in N$ (grouped by distances) against each other for temperatures $T = 0, \cdots, 5$, we get Figure (5.13) which verifies the previous equation as well as equation (5.29) by showing that $I \approx \Gamma^2$.

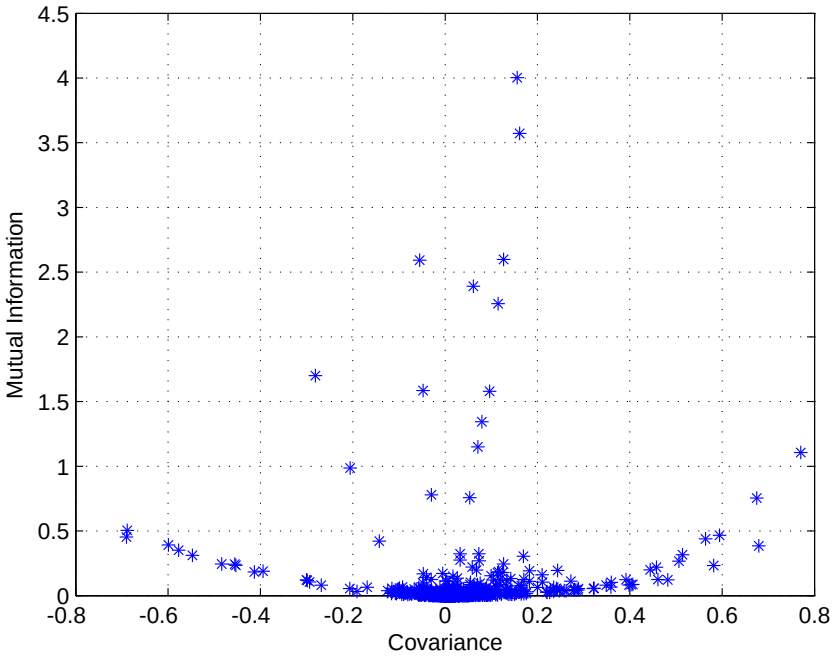

Figure 5.13: Mutual Information I versus covariance Γ values displaying the $I \approx \Gamma^2$ relationship

Chapter Summary

The Ising model is a system that displays phase transition (denoted by crossover temperature in the simulations). The effect of distance is evident for all the measures tested on the model and the conditional Mutual Information seem to able to detect indirect dependence of the sites due to nearest neighbour interaction, however the Mutual Information and time delayed Mutual Information as well as the Transfer Entropy does not yield a direction or any indication of 'causality'. We suspect that this is because Ising model is intrinsically symmetrical and thus the interactions are more or less equal in all directions. This is the fact which was used to obtain the formula that relates covariance and Mutual Information. Therefore, in the next chapter we discuss how to break the symmetry and how we 'amend' the Ising model to incorporate our idea.

Chapter 6

The amended Ising model

The amendment that shall be made on the well established Ising model in this chapter is meant to incorporate ‘causality’ on the model. We will begin by discussing different attempts to break the symmetry and how the amendment came about. Next we take a closer look at how we altered the Metropolis Monte Carlo (MMC) algorithm so that some transition probabilities are altered. We then simulated the amended Ising model and evaluated the results obtained by contrasting it with the results from the previous chapter. The amended model enables us to demonstrate that Transfer Entropy has the capability to detect the direction of ‘causality’ and furthermore identify the actual causal lag.

6.1 Replicating ‘causality’

In order to replicate ‘causality’ on the model, we need elements of time and dependency. Moreover since ‘causality’ is asymmetrical by nature, something needs to be done about the symmetrical nature of the Ising model. We tried a few alterations to break the symmetry, in order to see whether this will effect the values of the measures applied on the model and the relationship between them. We would like to tip the balance of some sites and create some sort of artificial ‘causality’ on the model so that we can verify the measures that claim to be able to detect ‘causality’.

6.1.1 Attempts and ideas

In our quest to observe the effects of 'causality', first we tried fixing the value of a site X on the lattice i.e $p_{s_X}(1) = 1$ and $p_{s_X}(-1) = 0$ to see the ripple effect it has onto the other sites. Unfortunately the values of covariance, Mutual Information and Transfer Entropy of any other site with the site X would be 0, since by fixing the value, s_X becomes a constant and $p_{s_X s_Y}(\alpha, \beta) = p_{s_X s_Y}(1, \beta) = p_{s_X}(\beta), \forall \alpha, \beta \in \{-1, 1\}$ for any site $Y \neq X$ on the lattice. Thus $p_{s_X s_Y}(\alpha, \beta) = p_{s_X}(\alpha) p_{s_Y}(\beta), \forall \alpha, \beta \in \{-1, 1\}$ and X is independent of any other site Y in the lattice. It becomes obvious when the covariance is written this way:

$$\Gamma(s_X, s_Y) = E(s_X s_Y) - E(s_X)E(s_Y) = s_X E(s_Y) - s_X E(s_Y) = 0. \tag{6.1}$$

Therefore, we proceed to other alterations.

Secondly, we tried to create a dependence of sites A and B on the site G in the lattice by equating the spins of the sites. We did this by interfering in the MMC algorithm so that when the site G is chosen for flipping consideration, $s_A = s_B = s_G$ is imposed with probability $1 - p_G$. And with probability $1 - p_G$ the normal MMC algorithm flipping consideration applies (just like if all the other sites was chosen). However as we later found out, if $s_A = s_B = s_G$ is imposed when G is selected then one would have problems when it comes to getting the probabilities, due to the fact that we do not know the order in which A, B and G is selected for flipping considerations. The estimated values of $P((s_A)_{n-1} \neq (s_A)_n)$ and $P((s_B)_{n-1} \neq (s_B)_n)$ will be different for each order.

To remedy this, we decided to change the mechanism so that the probabilities can be understood better. Instead of interfering when site G is chosen for flipping consideration, we decided to interfere when A and B is selected. The normal MMC process is maintained except when A and B is chosen for flipping consideration, when either of this happens we look at the value of s_G at the last sampled time (regardless of whether s_G has changed or not from the last sampling value) and if $s_G = 1$ we let the site be considered as usual but if $s_G = -1$ we do not allow any changes to the current state of the selected site. Notice that the probabilities of s_G changing is just like any other site (excluding A and B) on the lattice and is not interfered with. This is the algorithm that will shape the amended Ising model.

Essentially what this amendment does is make the changes of s_A and s_B dependent on s_G. In a way this gives G some control over the state of A and B, however it must be pointed out that this is not complete control due to the fact that when given permission to change by G, A and B can still chose not to change (since the change is done with probability γ_B). The condition (or the amendment) works in such a way that $(s_G)_{n-1}$ limits the ability of $(s_A)_{n-1}$ and $(s_B)_{n-1}$ to change so that the transition probabilities $P((s_A)_{n-1} \neq (s_A)_n)$ and $P((s_B)_{n-1} \neq (s_B)_n)$ will be altered. It must be pointed out the algorithm does not dictate what the states of A and B are supposed to be or whether it should be similar to G. Note that in our definition of the amended Ising model we have determined that the condition for letting s_A and s_B be considered for flipping consideration (each time it is randomly selected) is that $(s_G)_{n-1} = 1$. One could chose the condition to be $(s_G)_{n-1} = -1$ and the outcome would remain unchanged.

6.1.2 The Generating Mechanism

The two dimensional amended Ising model is generated using the standard MMC algorithm [62, 77, 27] as outlined in subsection (5.2.1) albeit interference whenever site A or B is chosen for flipping consideration. More formally, we will generate the amended Ising model using the algorithm outlined as follows. At each step in the algorithm a site chosen at random will be considered for flipping with a certain probability γ_B in equation (5.4). This apply for all sites except when A or B is selected. When this happens we look at the value of s_G at the last sampled time and if $s_G = 1$ we let the site be considered for flipping with probability γ_B as usual, however if $s_G = -1$, no change is allowed. Thus only one state of G ($s_G = 1$ in this case) allows sites A and B to be considered for flipping and therefore actually change. Hence, one can say that in this way any changes of A and B depends on G.

Let X be any site on the lattice not affected by our imposed condition. Then the transition probabilities (from one sample to the next) of s_X is approximately

$$P((s_X)_n = \alpha | (s_X)_{n-1} = \beta) = \begin{cases} 1 - \gamma_B, & \text{if } \alpha = \beta \text{ for any } \alpha, \beta \in \{-1, 1\} \\ \gamma_B, & \text{if } \alpha \neq \beta \text{ for any } \alpha, \beta \in \{-1, 1\} \end{cases}$$

where γ_B is the probability in equation (5.4). With the amendment, the values of $(s_A)_n$ and $(s_B)_n$ depends on $(s_G)_{n-1}$. For any $\alpha, \beta \in \{-1, 1\}$ the transition probability of s_A can be written as

$$P((s_A)_n = \alpha | (s_A)_{n-1} = \beta) = \begin{cases} 1 - \gamma_B P((s_G)_{n-1} = 1) = 1 - \gamma_B p_{s_G}(1), & \text{if } \alpha = \beta \\ \gamma_B P((s_G)_{n-1} = 1) = \gamma_B p_{s_G}(1), & \text{if } \alpha \neq \beta, \end{cases}$$

and $P((s_B)_n = \alpha | (s_B)_{n-1} = \beta) = P((s_A)_n = \alpha | (s_A)_{n-1} = \beta)$. We denote $p_{s_G}(1) = P(s_G = 1) = P((s_G)_{n-1} = 1)$ since the marginal probabilities are the same for any n as we are taking the time average as discussed in subsection (5.2.2). Note that s_G is not altered in any way and should be treated just like any other unamended site on the lattice. Due to the nearest neighbour nature of the Ising model, the interactions between neighbours are accounted for through the Hamiltonian $\mathcal{H}$ in γ_B, therefore the neighbours of A and B might also have their transition probabilities altered.

If a certain site Y does not affect another site X,

$$P((s_X)_n = \alpha | (s_X)_{n-1} = \beta, (s_Y)_{n-1} = \gamma) = P((s_X)_n = \alpha | (s_X)_{n-1} = \beta)$$

for any $\alpha, \beta, \gamma \in \{-1, 1\}$. However, due to our amendment, this not true for probabilities $P((s_A)_n = \alpha | (s_A)_{n-1} = \beta, (s_G)_{n-1} = \gamma)$ and $P((s_B)_n = \alpha | (s_B)_{n-1} = \beta, (s_G)_{n-1} = \gamma)$. Define $Q^{(\tau)}_{sgn(\gamma)}$ such that

$$Q^{(\tau)}_{sgn(\gamma)} = P(\text{condition fulfilled} | (s_G)_{n-\tau} = \gamma) = P((s_G)_{n-1} = 1 | (s_G)_{n-\tau} = \gamma) \tag{6.2}$$

where

$$sgn(\gamma) = \begin{cases} + & \text{if } \gamma = 1 \\ - & \text{if } \gamma = -1. \end{cases}$$

With this, for any $\alpha, \beta, \gamma \in \{-1, 1\}$ we get that

$$P((s_A)_n = \alpha | (s_A)_{n-1} = \beta, (s_G)_{n-1} = \gamma) = \begin{cases} 1 - \gamma_B Q^{(1)}_{sgn(\gamma)}, & \text{if } \alpha = \beta \\ \gamma_B Q^{(1)}_{sgn(\gamma)}, & \text{if } \alpha \neq \beta. \end{cases}$$

Therefore the corresponding ratios become

$$\frac{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta,(s_G)_{n-1}=\gamma)}{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta)}=\begin{cases}\frac{1-\gamma_B Q^{(1)}_{sgn(\gamma)}}{1-\gamma_B p_{s_G}(1)}, & \text{if } \alpha=\beta\\ \frac{\gamma_B Q^{(1)}_{sgn(\gamma)}}{\gamma_B p_{s_G}(1)}=\frac{Q^{(1)}_{sgn(\gamma)}}{p_{s_G}(1)}, & \text{if } \alpha\neq\beta,\end{cases}$$

and the values differ for $\gamma=1$ and $\gamma=-1$. The same applies for site B. This ratio is exactly what the Transfer Entropy $T^{(1)}_{GA}$ and $T^{(1)}_{GB}$ as in equation (5.18) takes into account and sums up. To illustrate how this works, we highlight the fact that if $(s_G)_{n-1}=1$ i.e $Q^{(1)}_{+}=P((s_G)_{n-1}=1|(s_G)_{n-1}=1)=1$, we obtain

$$\frac{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta,(s_G)_{n-1}=1)}{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta)}=\begin{cases}\frac{1-\gamma_B Q^{(1)}_{+}}{1-\gamma_B p_{s_G}(1)}=\frac{1-\gamma_B}{1-\gamma_B p_{s_G}(1)}, & \text{if } \alpha=\beta\\ \frac{Q^{(1)}_{+}}{p_{s_G}(1)}=\frac{1}{p_{s_G}(1)}, & \text{if } \alpha\neq\beta,\end{cases}$$

and when $(s_G)_{n-1}=-1$, we get $Q^{(1)}_{-}=P((s_G)_{n-1}=1|(s_G)_{n-1}=-1)=0$ so that

$$\frac{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta,(s_G)_{n-1}=-1)}{P((s_A)_n=\alpha|(s_A)_{n-1}=\beta)}=\begin{cases}\frac{1-\gamma_B Q^{(1)}_{-}}{1-\gamma_B p_{s_G}(1)}=\frac{1}{1-\gamma_B p_{s_G}(1)}, & \text{if } \alpha=\beta\\ \frac{Q^{(1)}_{-}}{p_{s_G}(1)}=0, & \text{if } \alpha\neq\beta.\end{cases}$$

6.1.3 Incorporating causal lags

We can generalize the 'dependency' to be at a chosen causal lag t_G by imposing that A and B can only change states at time n (i.e having different states than at time $n-1$) if the state of G is equal to 1 at time step $n-t_G$. Now the condition is set to be $(s_G)_{n-t_G}=1$ instead of $(s_G)_{n-1}=1$. For sites A and B the transition probabilities of their states will be

$$P((s_A)_n=\alpha|(s_A)_{n-t_G}=\beta)=\begin{cases}1-\gamma_B P((s_G)_{n-t_G}=1)=1-\gamma_B p_{s_G}(1), & \text{if } \alpha=\beta\\ \gamma_B P((s_G)_{n-t_G}=1)=\gamma_B p_{s_G}(1), & \text{if } \alpha\neq\beta,\end{cases}$$

where $t_G=1$ is a special case explained in the previous subsection. Once again, we denote $p_{s_G}(1)=P(s_G=1)=P((s_G)_{n-t_G}=1)$ since the marginal probabilities are the same for any $n-t_G$ due to the utilization of time average approximation as discussed in subsection

(5.2.2). With the condition $(s_G)_{n-t_G} = 1$ the value of Q can be written as

$$Q^{(\tau)}_{sgn(\gamma)} = P(\text{condition fulfilled}|(s_G)_{n-\tau} = \gamma) = P((s_G)_{n-t_G} = 1|(s_G)_{n-\tau} = \gamma).$$

Therefore the ratios in Transfer Entropy $T^{(\tau)}_{GA}$ as in equation (5.18) will be

$$\frac{P((s_A)_n = \alpha|(s_A)_{n-1} = \beta, (s_G)_{n-\tau} = \gamma)}{P((s_A)_n = \alpha|(s_A)_{n-1} = \beta)} = \begin{cases} \frac{1-\gamma_B Q^{(\tau)}_{sgn(\gamma)}}{1-\gamma_B p_{s_G}(1)}, & \text{if } \alpha = \beta \\ \frac{\gamma_B Q^{(\tau)}_{sgn(\gamma)}}{\gamma_B p_{s_G}(1)} = \frac{Q^{(\tau)}_{sgn(\gamma)}}{p_{s_G}(1)}, & \text{if } \alpha \neq \beta. \end{cases}$$

for any $\alpha, \beta, \gamma \in \{-1, 1\}$. The value of $Q^{(\tau)}_{sgn(\gamma)}$ changes for different τ and this is the heart of the Transfer Entropy value. The changes of $Q^{(\tau)}_{sgn(\gamma)}$ enables us to use the Transfer Entropy to detect the exact causal lag t_G by comparing different values of Transfer Entropy with different τ values. $T^{(\tau)}_{GA}$ should be the largest when $\tau = t_G$ as $Q^{(t_G)}_{sgn(\gamma)}$ is either 1 or 0.

Again, for all the simulations the interaction strength is set to be $J = 1$ and the Boltzmann constant is fixed as $K_B = 1$ for all the simulations. As in the Ising model, we let the system run up to 2000 samples before sampling at every $N = L^2$ time steps and this is done for more than 100 temperature values T ranging from 0 to 5. For illustration purposes, $L = 10$ is usually utilized unless stated otherwise. The simulations displayed in this chapter will be displaying values of the amended Ising model with periodic boundary conditions and $\mathcal{T} = 100000$ samples for each site.

6.2 Measures on the amended Ising model

The formulas that we have defined in subsection (5.3) apply here as well. We would like to point out that it is the probabilities that change not the formulas. Recall that A, B and G are sites on the lattice at coordinates $[1, 1]$, $[2, 2]$ and $[3, 3]$ respectively. The amendment on the model are intended to make changes of s_A and s_B dependent on s_G.

6.2.1 Observables for verification of the critical point

The amended Ising model also generates a crossover temperature T_c where most measures peak due to the fully connected lattice which can be seen from values of magnetisation M in

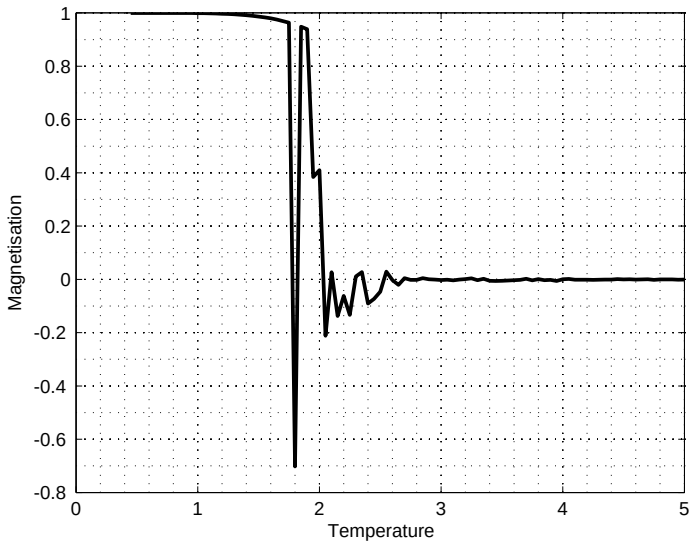

Figure 6.1: Values of magnetisation M using equation (5.13) on amended Ising model with $t_G = 1$ approaches 0 at T_c.

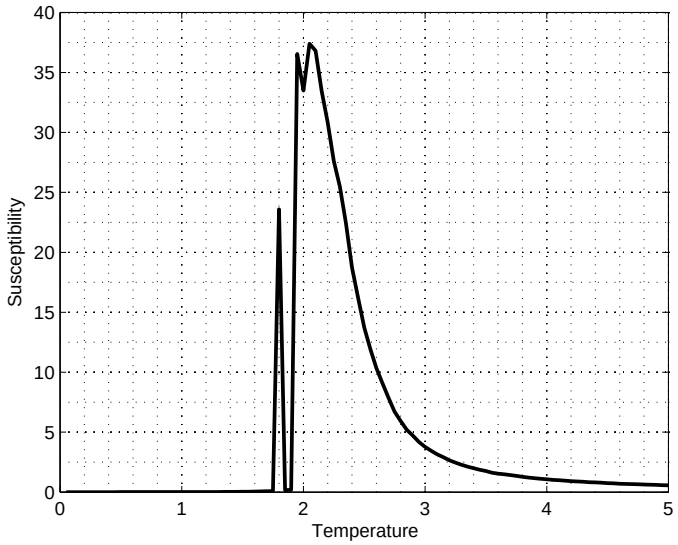

Figure 6.2: Values of susceptibility χ using equation (5.14) on amended Ising model with $t_G = 1$ peaks at T_c.

Figure (6.1), susceptibility χ in Figure (6.2) and covariance as well as Mutual Information in Figure (6.3). We say most measures because, in Figure (6.4) we see that this does not apply to Transfer Entropy for one of the directions. The Transfer Entropy of G to A at time

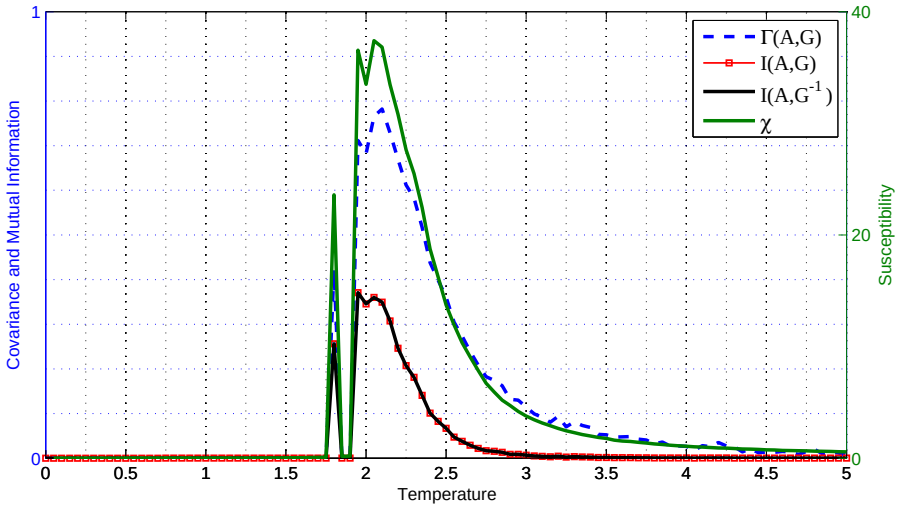

Figure 6.3: Values of covariance $\Gamma(A, G)$ using equation (5.10), Mutual Information $I(A, G)$ using equation (5.15), time delayed Mutual Information $I(A, G^{-1})$ using equation (5.17) and susceptibility χ using equation (5.14) across temperature T on amended Ising model with $t_G = 1$.

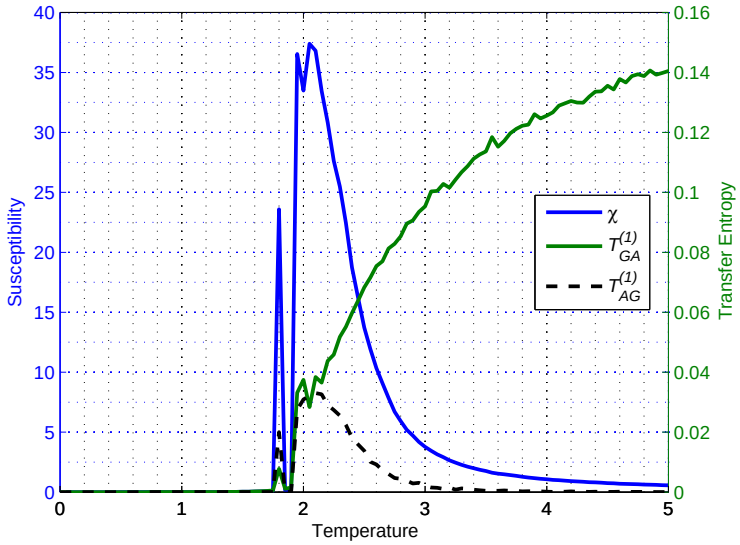

Figure 6.4: Values of susceptibility χ using equation (5.14) as well as Transfer Entropy $T^{(1)}_{GA}$ and $T^{(1)}_{AG}$ both using equation (5.18) across temperature T on amended Ising model with $t_G = 1$. $T^{(1)}_{GA}$ does not decrease to 0 after T_c which indicates that G causes A.

1, $T_{GA}^{(1)}$ does not decrease to 0 for larger temperatures after T_c. This difference coincides with the condition that we have imposed on the model. This only happens when Transfer Entropy is applied on the correct direction and exact causal lag t_G imposed on the model. One can see in Figure (6.3) that once again (as was the case on the unamended Ising model) there is almost no difference between $I(A, G)$ and $I(A, G^{-1})$, even though the causal lag was imposed at $t_G = 1$. From Figures (6.1) and (6.2) we will use $T_c = 2.1$ in obtaining the value of reduced temperature $\frac{T-T_c}{T_c}$ for figures in this chapter resulting from $L = 10$ simulation of the amended Ising model with $t_G = 1$.

6.2.2 The influence of distance

In subsection (5.3.3), we have seen that on the unamended Ising model, distance is the main factor that influences the strength of the different measures between sites. We observe the

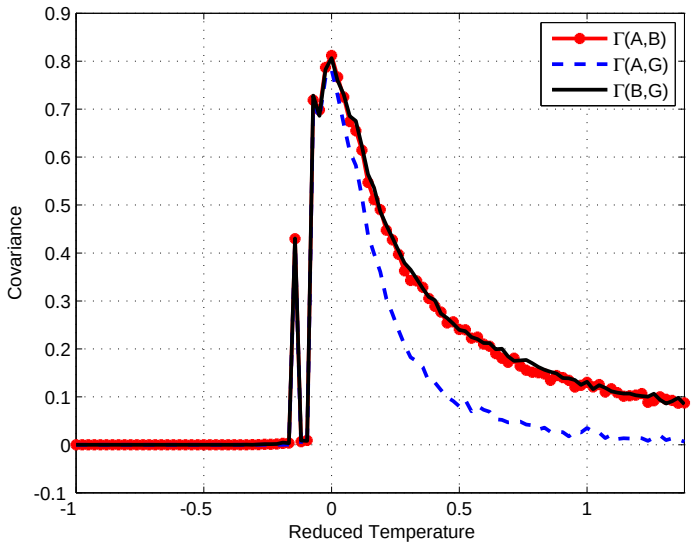

Figure 6.5: Covariance $\Gamma(A, B)$, $\Gamma(A, G)$ and $\Gamma(B, G)$ using equation (5.10) in amended Ising model with $t_G = 1$. $\Gamma(A, G) < \Gamma(A, B) \approx \Gamma(B, G)$ due to distance.

same behaviour in Figures (6.5), (6.6) and (6.7) where the covariance, Mutual Information and the time delayed Mutual Information between the three sites A, B and G are plotted.

Recall that for the amended model, we make A and B dependent on G, however this does not change the fact that B is situated between A and G. It seems that these measure

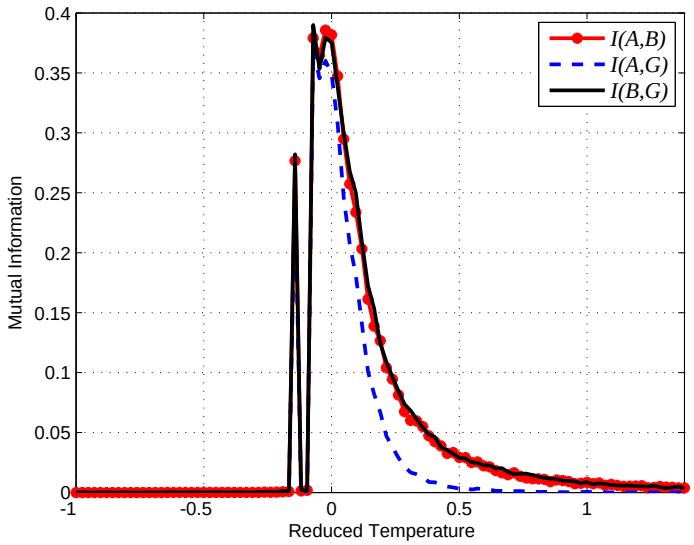

Figure 6.6: Mutual Information $I(A,B), I(A,G)$ and $I(B,G)$ using equation (5.15) in amended Ising model with $t_G = 1$. $I(A,G) < I(A,B) \approx I(B,G)$ due to distance.

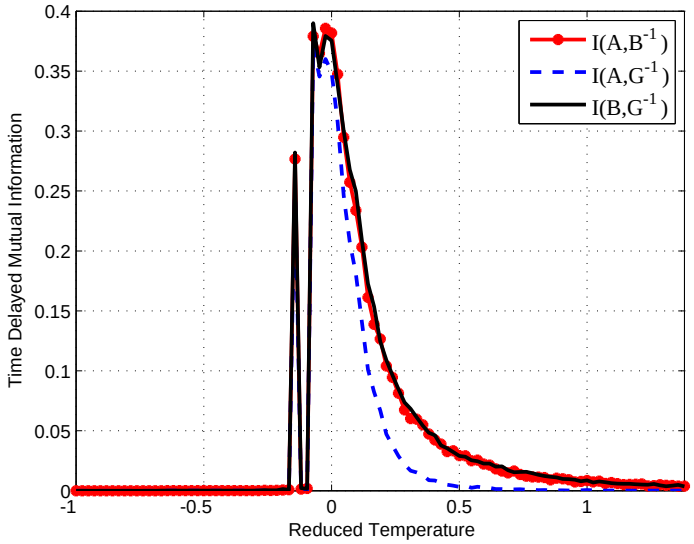

Figure 6.7: Time delayed Mutual Information $I(A,B^{-1})$, $I(A,G^{-1})$ and $I(B,G^{-1})$ using equation (5.17) versus reduced temperature $\frac{T-T_c}{T_c}$ in amended Ising model with $t_G = 1$. $I(A,G^{-1}) < I(A,B^{-1}) \approx I(B,G^{-1})$ due to distance.

are oblivious to the imposed mechanism and their values reflect those on the unamended Ising model. From comparing Figure (6.6) and Figure (6.7) one can see that there are minimal differences between the Mutual Information and time delayed Mutual Information simulation outcomes. This is because time delayed Mutual Information only takes into account the static probabilities at different time steps (sampled time from simulation) and basically compares the state of the sites. Recall that what $(s_G)_{n-1}$ does is limit the ability of $(s_A)_{n-1}$ and $(s_B)_{n-1}$ to change but does not dictate that the states between the three sites will be identical therefore the average and join average values of s_A and s_B may very well remain unchanged although the transition probabilities were altered. Thus it seems that even time delayed Mutual Information cannot detect the imposed mechanism.

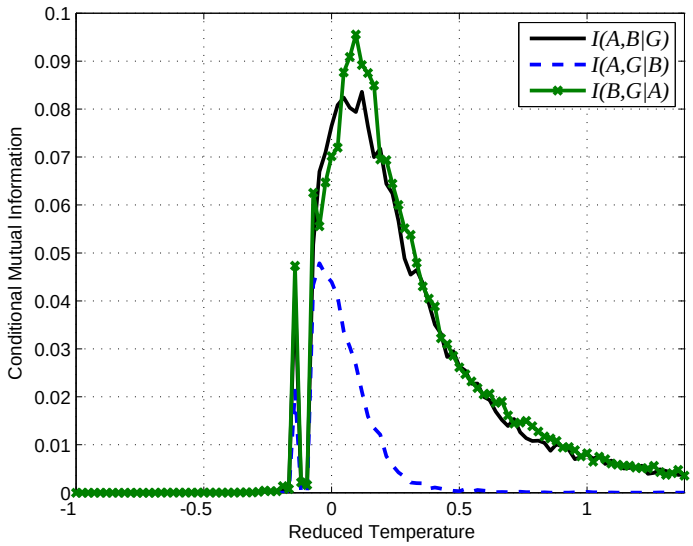

Figure 6.8: $I(A, B|G)$, $I(A, G|B)$ and $I(B, G|A)$ equation (5.16) versus $\frac{T-T_c}{T_c}$ for $t_G = 1$

Figure (6.8) is in agreement with Figure (5.7) of the unamended Ising model where both shows quite distinctly that $I(A, G|B)$ gives the lowest conditional Mutual Information value when the values of $I(A, B|G)$, $I(A, G|B)$ and $I(B, G|A)$ are compared. The fact that we have imposed the condition so that A and B depend on G seems to make no difference. Again this is due to the fact that B is situated between A and G on the lattice and this is what both time delayed and conditional Mutual Information detect rather than the implanted dependency. If one says that conditional Mutual Information in Figure (6.8) indicate that

B causes the relationship between A and G, this is a contradiction since we clearly have set the model so that G causes A as well as B. Therefore we conclude that conditional Mutual Information without time delays where transition probabilities are not taken into account will not be useful in detecting our definition of 'causality' and the imposed mechanism. We move on to time delayed version of conditional Mutual Information which is better known as Transfer Entropy.

6.2.3 Measures on $L = 25$

From most of the other figures in this chapter, one can clearly see that the peaks at T_c is not as clear cut as it was the unamended Ising model. There seems to be an initial lower peak before the actual peak at T_c in the $L = 10$ lattice with sample size of $\mathcal{T} = 100000$ for any t_G that is used. We claim that this is just a fluctuation due to the small length of the lattice and the general behaviour on the lattice is not affected by this. We shall illustrate by displaying values on lattice with length $L = 25$ alongside the $L = 10$ analogous to subsection (5.3.4).

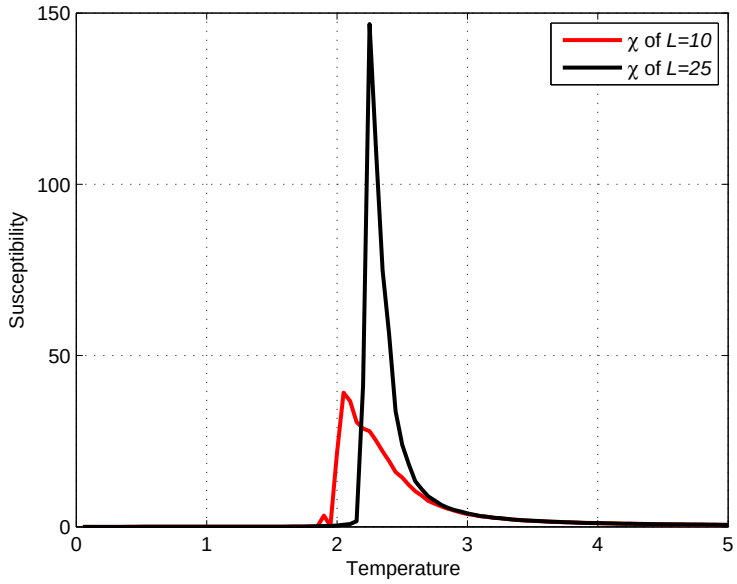

Figure 6.9: Values of susceptibility χ in equation (5.14) across temperature T on amended Ising model with $t_G = 10$ for $L = 10, 25$.

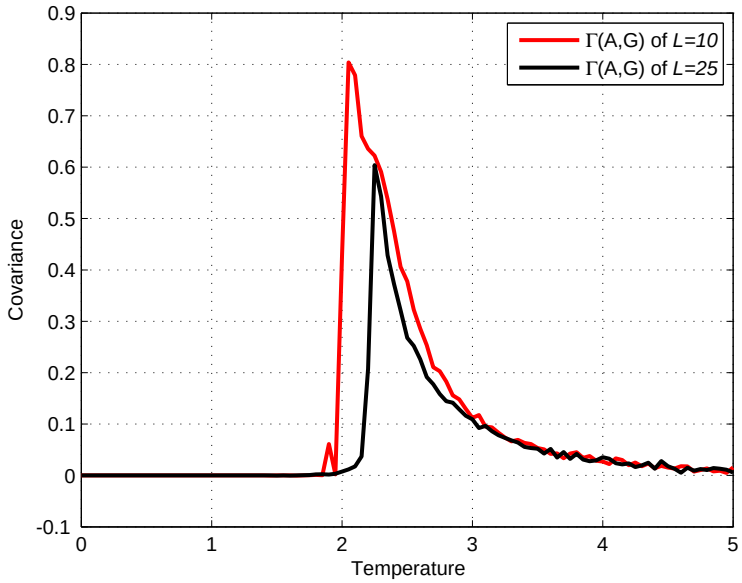

Figure 6.10: Values of covariance $\Gamma(A, G)$ using equation (5.10) across temperature T on amended Ising model with $t_G = 10$ for $L = 10, 25$.

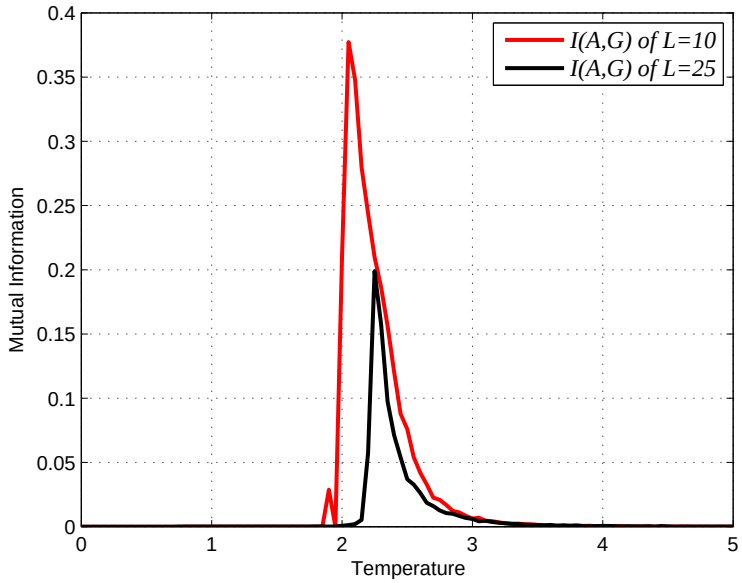

Figure 6.11: Values of Mutual Information $I(A, G)$ using equation (5.15) on amended Ising model with $t_G = 10$ for varying values of L.

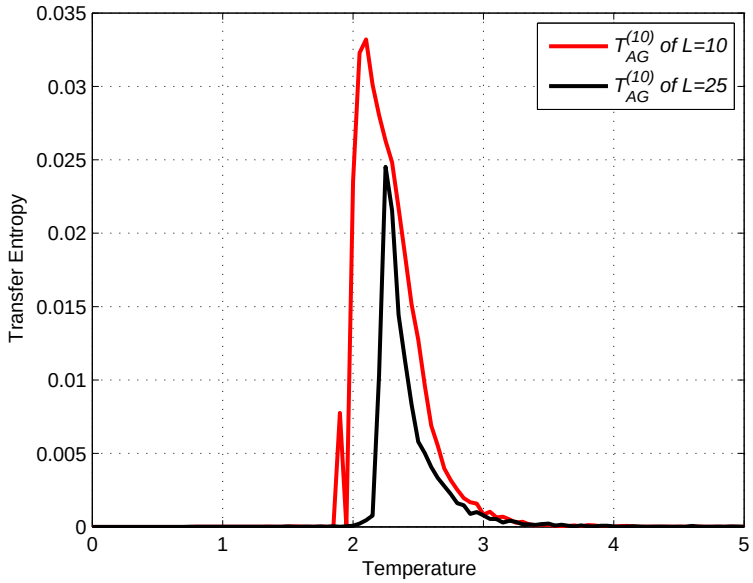

Figure 6.12: Values of $T_{AG}^{(10)}$ using equation (5.18) across temperature T on amended Ising model with $t_G = 10$ for $L = 10, 25$.

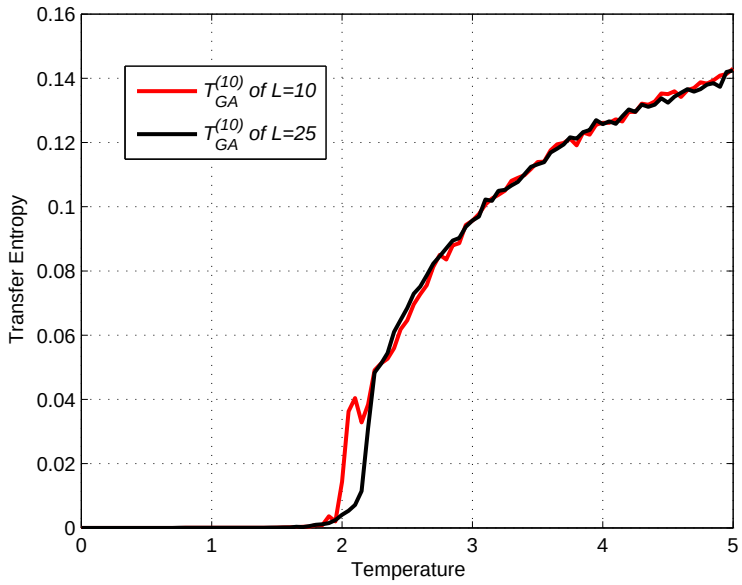

Figure 6.13: Values of $T_{AG}^{(10)}$ using equation (5.18) across temperature T on amended Ising model with $t_G = 10$ for $L = 10, 25$. Both indicates that G causes A.

Figures (6.9), (6.10), (6.11), (6.12) and (6.13) display the values of susceptibility χ, covariance $\Gamma(A,G)$, Mutual Information $I(A,G)$ and Transfer Entropy $T_{AG}^{(10)}$ as well as $T_{GA}^{(10)}$ on the amended Ising model with $t_G = 10$ for lattice lengths of $L = 10, 25$ so that $N = 100, 625$. The peaks clearly show that T_c does indeed exist in our model. In Figure (5.9) we observe that the value of χ increases as L increases since $\chi \to \infty$ as $L \to \infty$. The crossover temperature of $L = 25$ is $T_c \approx 2.25$ which is closer to the real T_c. As in subsection (5.3.4), the figures indicate sharper and more precise detection of T_c in the $L = 25$ lattice. We reiterate our suspicion that this is due to much bigger lattice of $L = 25$ where the influence of the two sites on each other is weaker than in the $L = 10$ lattice. The fact that the measures attained maximum values near T_c is consistent to our observation on $L = 10$. Moreover, Figure (6.13) that displays the behaviour of Transfer Entropy in indicating direction of G causes A does not seem to be affected by the lattice sizes.

6.3 Transfer Entropy results

Figure (6.4) showed that Transfer Entropy indicating causal direction of $G \to A$ at the implanted causal lag $t_G = 1$, as the values of $T_{AG}^{(1)}$ go down to zero after T_c. This is also the

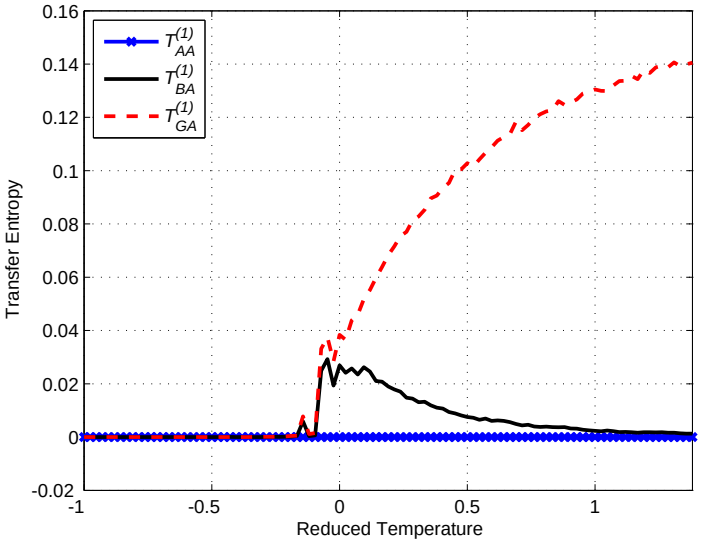

Figure 6.14: $T_{AA}^{(1)}$, $T_{BA}^{(1)}$ and $T_{GA}^{(1)}$ using equation (5.18) versus reduced temperature $\frac{T-T_c}{T_c}$ on the amended Ising model with $t_G = 1$. Clearly $G \to A$ at $\tau = 1$.

case when $T^{(1)}_{BG}$ is compared to $T^{(1)}_{GB}$. This is something we do not see for the unamended Ising model where all values of Transfer Entropy peaks at T_c and then goes down to 0 for higher temperatures. In Figure (6.14), we see that Transfer Entropy of site A is illustrated. Contrast this to Figure (5.8) where the exact same values were plotted on the unamended Ising model. Firstly we see that $T^{(1)}_{AA}$ is zero as expected, but more importantly we see that $T^{(1)}_{GA}$ is very different from $T^{(1)}_{BA}$ and this clearly indicates that G causes A at $\tau = 1$ and B does not. Values like $T^{(1)}_{AG}$ in Figure (6.4) and $T^{(1)}_{BA}$ in Figure (6.14) peak at T_c and then they reduce to 0 at higher temperature. We suspect that this is due to the fact that at T_c the whole lattice is strongly correlated thus there is no clear direction in which 'causality' may occur. From a different point of view, one could say that any site may equally likely influence or 'cause' any other sites hence we have that the Transfer Entropy $T^{(\tau)}_{XY}$ peaks at T_c whenever $X \neq G$ and $X \neq Y$.

6.3.1 Transfer Entropy as a causal lag indicator

In Figure (6.15), values of $\tau = 1, 2, 3$ of $T^{(\tau)}_{GA}$ are plotted. The first plot of $T^{(1)}_{GA}$ is exactly the same as $T^{(1)}_{GA}$ in Figures (6.4) and (6.14). Figure (6.15) shows that for values other than

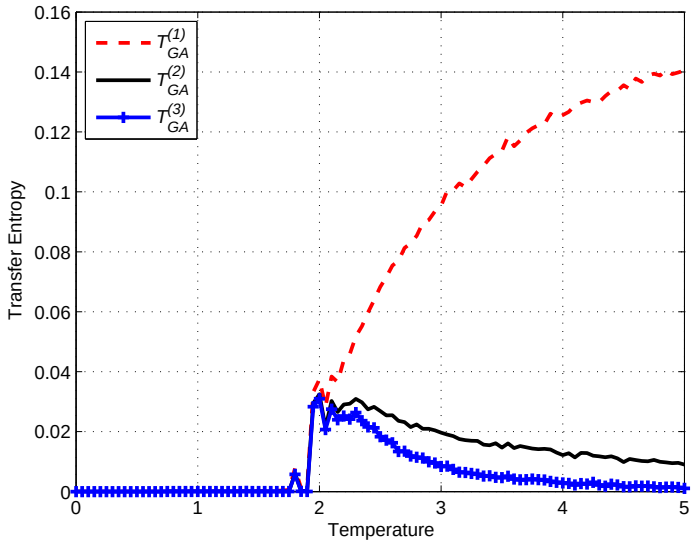

Figure 6.15: $T^{(1)}_{GA}$, $T^{(2)}_{GA}$ and $T^{(3)}_{GA}$ using equation (5.18) versus T for amended Ising model with $t_G = 1$. $T^{(1)}_{GA}$ indicates $G \to A$ at $\tau = 1$.

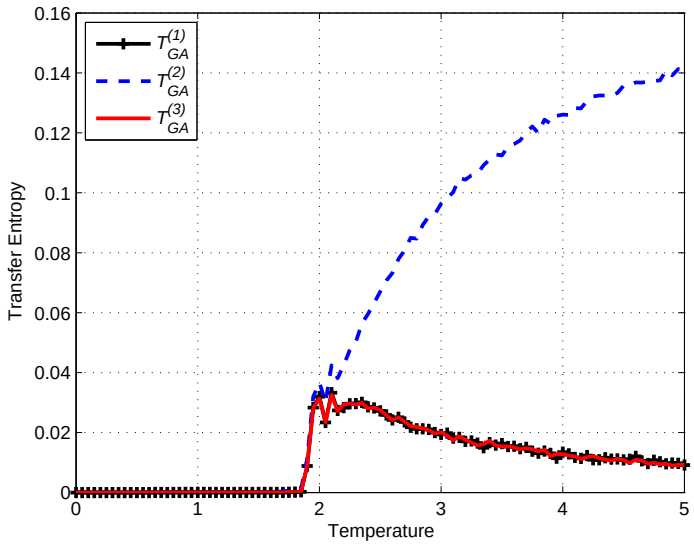

Figure 6.16: $T^{(1)}_{GA}$, $T^{(2)}_{GA}$ and $T^{(3)}_{GA}$ using equation (5.18) versus T for amended Ising model with $t_G = 2$. $T^{(2)}_{GA}$ correctly indicates $G \to A$ at $\tau = 2$.

$t_G = 1$, the Transfer Entropy $T^{(\tau)}_{GA}$ eventually goes to 0 after T_c. Therefore, in Figure (6.15) the Transfer Entropy correctly indicates that $t_G = 1$ in the model. However, the value of $T^{(2)}_{GA} > T^{(3)}_{GA}$ since the transition probability $P((s_A)_n|(s_A)_{n-2})$ in $T^{(2)}_{GA}$ is effected more than transition probability $P((s_A)_n|(s_A)_{n-3})$ in $T^{(3)}_{GA}$ by the changes imposed by $t_G = 1$.

This is also manifested in Figure (6.16) where the causal lag is set to be $t_G = 2$ in the amended model and $T^{(\tau)}_{GA}$ is calculated for $\tau = 1, 2, 3$. We see that it clearly shows that $T^{(2)}_{GA}$ has the highest value so that the detected causal lag is correctly $t_G = 2$. Moreover the values of $T^{(1)}_{GA}$ and $T^{(3)}_{GA}$ are almost identical, due to the fact that $\tau = 1$ and $\tau = 3$ are the same distance away from $t_G = 2$ so that the transition probability $P((s_A)_n|(s_A)_{n-1})$ and transition probability $P((s_A)_n|(s_A)_{n-3})$ are equally affected. This effect of distance from the predetermined causal lag t_G can be clearly seen in Figure (6.17) where we use $T^{(\tau)}_{GA}$ on the amended Ising model with $t_G = 10$. We have plotted Transfer Entropy values for $\tau = 6$ to $\tau = 10$, clearly illustrating that $T^{(10)}_{GA}$ gives the highest value thus indicating that $t_G = 10$. The rest of the Transfer Entropy values reduces to 0 but at different rates depending on the distance of τ from t_G. The further away from t_G, the faster it decreases to 0. We will discuss more about the relationship of distance of τ from t_G in relation with transition probabilities

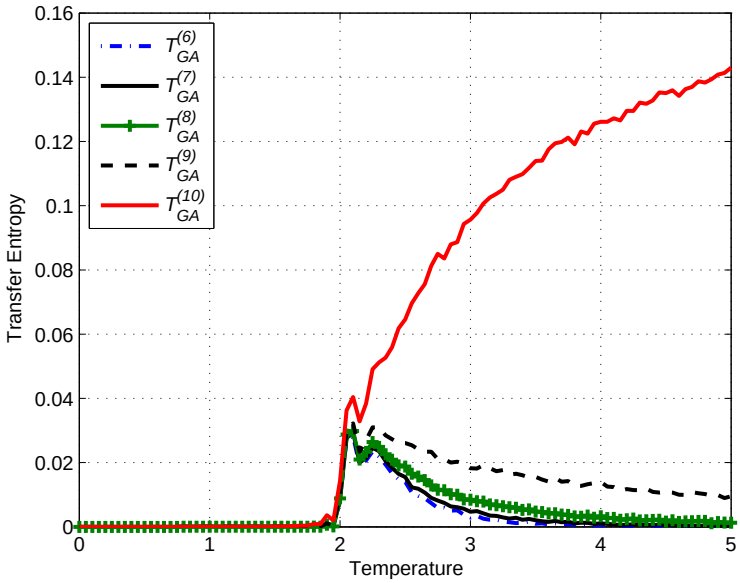

Figure 6.17: $T_{GA}^{(6)}$, $T_{GA}^{(7)}$, $T_{GA}^{(8)}$, $T_{GA}^{(9)}$ and $T_{GA}^{(10)}$ using equation (5.18) versus T for amended Ising model with $t_G = 10$. $T_{GA}^{(10)}$ indicates $G \rightarrow A$ at $\tau = 10$.

and its effect on the value of $Q_{sgn(\gamma)}^{(\tau)}$ in subsection (7.2.1). Nevertheless, the clear difference between $T_{GA}^{(\tau)}, \tau \neq t_G$ and $T_{GA}^{(t_G)}$ that differentiates t_G from the other τ essentially identifies the causal lag.

6.3.2 Discussions on the nature of Transfer Entropy

We have seen that covariance, Mutual Information, conditional Mutual Information and even time delayed Mutual Information have failed to detect the amendment we have made on the Ising model. This is mostly due to the fact that the amendment on the model effects the transition probabilities and not the static probabilities of the states of the sites. The transition probability quantifies the possible changes that can occur in a system and change is what happens in 'causality'.

The static probabilities being the Boltzmann distribution influenced only by nearest neighbour interactions are manifested in these other measures. Used individually on each site, the measures indicate the distance of the sites from each other which is logical on a lattice where nearest neighbour interaction is the main interaction. In addition to detecting

the implanted changes, the Transfer Entropy also takes into account the distances in terms of the amplitude of the measure. We have seen an example of this in Figure (5.8) on the unamended Ising model for Transfer Entropy values that peak at T_c.

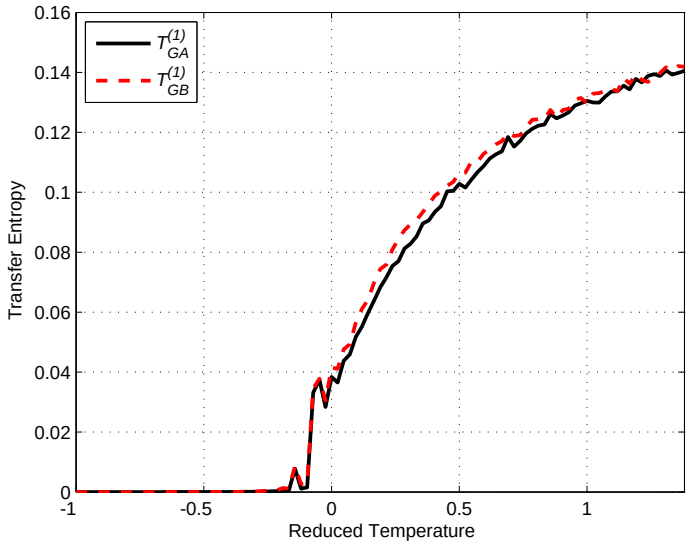

Figure 6.18: Transfer Entropy $T^{(1)}_{GA}$ and $T^{(1)}_{GB}$ using equation (5.18) for the unamended Ising model with $t_G = 1$. Generally $T^{(1)}_{GB} > T^{(1)}_{GA}$ due to distance on the lattice.

Another example on Transfer Entropy values that actually detect the causal lag is given in Figure (6.18) where $T^{(1)}_{GB}$ is mostly larger than $T^{(1)}_{GA}$ since G is closer B than to A, although both values do not reduce to zero. We suspect that the reason these values increase after T_c is simply because of the nature of Boltzmann distribution where probability of each site getting selected for flipping consideration is approaching uniform for higher temperatures, therefore allowing our mechanism to be implemented much more frequently than at lower values. Figure (6.19) illustrates that the values seem to be stabilizing to a certain fixed value. In the next chapter, the simple model approximates the values of the amended Ising model at these higher temperatures.

It is worth mentioning again that the restriction and conditioning that is done on the model is to create a 'causal' relationship. We have seen that the interpretation of Transfer Entropy relates to detecting the change through transition probabilities and also in terms of influences apparent in predictions. However what we have done here is to define the

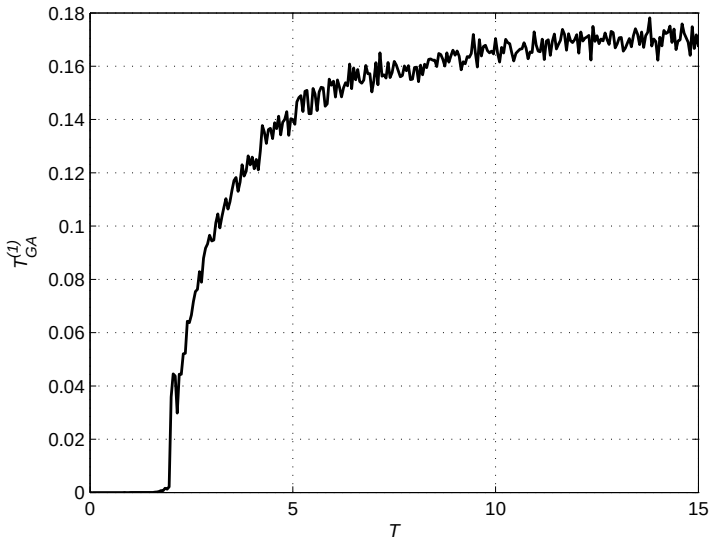

Figure 6.19: Transfer Entropy $T^{(1)}_{GA}$ for values up to temperature $T = 15$ using equation (5.18) on the amended Ising model with $t_G = 1$.

influence more as a restriction of one variable on another, in a way that a value of a variable will affect the possible changes of the other variable. It is this idea that we will continue to expand on the toy model in the coming chapters.

Chapter Summary

We have seen that the Transfer Entropy successfully indicated the direction of our artificially implanted 'causality' as well as the causal lag at which it was implanted. The values of conditional Mutual Information and time delayed Mutual Information gives more or less the same values as in the unamended Ising model and completely misses the amended part. From the figures one can see that like the other measures Transfer Entropy values peak at T_c and then reduces to 0 for most sites and direction where no 'causality' is detected. However when a 'causal' relationship is identified at the exact causal lag, the values of Transfer Entropy for that direction will keep on increasing even after T_c until the probability stabilizes at higher temperatures. We conclude that the Transfer Entropy is certainly worth focusing our attention on and thus proceed to investigate this measure for processes with higher number of states.

Chapter 7

A toy model

In an attempt to understand Transfer Entropy better, we apply it to a toy model where we can control the 'causal' connections. To incorporate a higher number of states, we decided to go back to basics and simply generate three random variables (in the form of stochastic processes) over a certain length of time. However, similar to the amended Ising model we restrict the changes for two of the variables and impose a condition to make to it dependent on another variable. We do this for three different cases of the general model. The challenge is to use Transfer Entropy to detect these 'causal' relationships and the exact causal lags.

7.1 A simple model

Assume we have a model of stochastic processes X, Y and Z that can assume values in the set of states $A = \{-1, 1\}$ at every time step $n = 1, \cdots, S$ where S is the length of stochastic process. Define X_n, Y_n and Z_n to be the values of process X, Y and Z at time step n respectively. Let μ_X, μ_Y and μ_Z be the independent (not influenced by other processes) probabilities that the variables X, Y and Z changes at every time step respectively. We supplement the dynamics by the special restriction on X and Y such that they are only allowed to do the stochastic swap with probability μ_X and μ_Y if the state of Z_{n-t_Z} fulfills a certain condition. For this simple model, we choose the condition to be $Z_{n-t_Z} = 1$. Without loss of generality, from here on we use $t_z = 1$ unless specified otherwise. To illustrate the mechanism, see Figure (7.1), where one can clearly see that

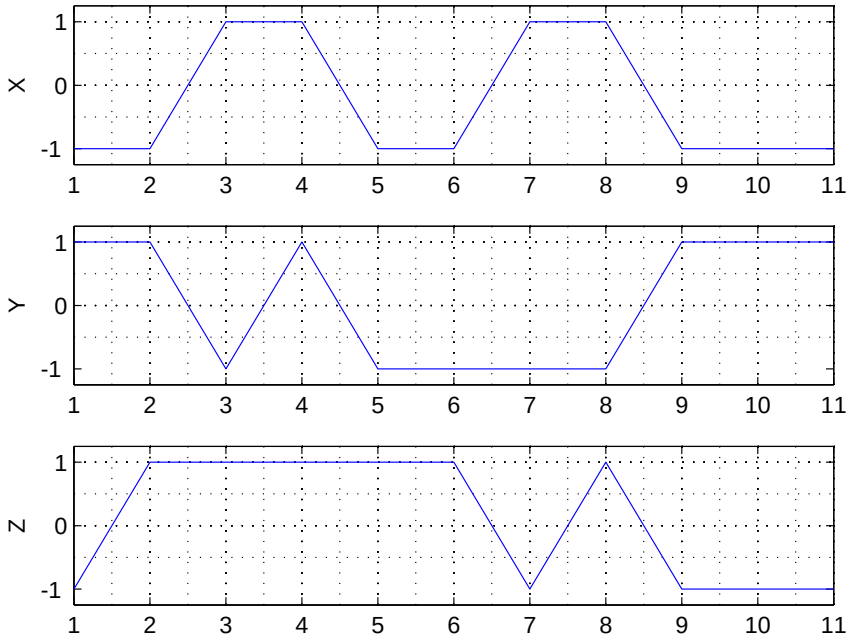

Figure 7.1: Simple model with $S = 11$ and $t_Z = 1$

when $Z_n = -1$ (at time steps $n = 1, 7, 9, 10, 11$), $X_{n+1} = X_n$ and $Y_{n+1} = Y_n$ since it is not allowed to change state. Figure (7.2) is just the same graph with more of time steps displayed for a clearer depiction.

In Figure (7.2) the processes were initialized randomly and independently, and this is not unlike the situation in the Ising model at higher temperatures due to the nature of Boltzmann distribution that tends to flatten out distributions for higher temperature. In higher temperature of the Ising model the distribution approaches uniformity, thus in a way the simple model is modelling the amended Ising model at higher temperatures. On the amended Ising model we had that $\mu_X = \mu_Y = \mu_Z = \gamma_B$ given by transition probability in equation (5.4). Therefore we expect to see that Transfer Entropy will clearly distinguish the direction as we have seen on the amended Ising model.

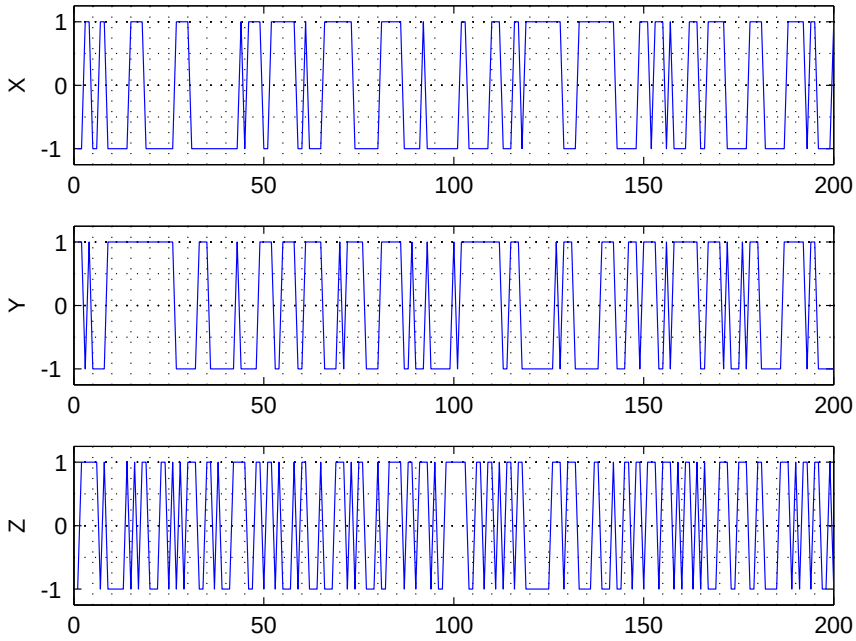

Figure 7.2: Simple model with $S = 200$ and $t_Z = 1$

7.1.1 Probabilities on the simple model

If the processes are initialized randomly and independently so that all initial probabilities are uniform i.e. $P(X_1 = -1) = P(X_1 = 1) = \cdots = P(Z_1 = 1) = \frac{1}{2}$, then for $\alpha \in A$,

$$\begin{aligned} P(X_2 = \alpha) &= P(X_1 = \alpha)P(X_2 = X_1) + P(X_1 = -\alpha)P(X_2 \neq X_1) \\ &= \frac{1}{2}(1 - P(X_2 \neq X_1)) + \frac{1}{2}P(X_2 \neq X_1) = \frac{1}{2}. \end{aligned}$$

One can apply this recursively so that $P(X_n = -1) = P(X_n = 1) = \frac{1}{2}$ for any n. The same applies for Y and Z. Therefore if the processes are initialized uniformly, the static probabilities do not depend on the transition probabilities and will always be $\frac{1}{2}$. Joint probabilities are the product of marginal probabilities. All joint probabilities of two processes becomes $\frac{1}{4}$. Joint probabilities of three processes becomes $\frac{1}{8}$ and so on. An example for $n = 1$ would be, $P(X_1 = \alpha, Z_1 = \beta) = P(X_1 = \alpha)P(Z_1 = \beta) = \frac{1}{4}$ for any $\alpha, \beta \in A$.

Consequently for $n = 2$, we have that for any $\alpha, \beta \in A = \{-1, 1\}$,

$$\begin{aligned}
P(X_2 = \alpha, Z_2 = \beta) = &P(X_1 = \alpha, Z_1 = \beta)P(X_2 = X_1)P(Z_2 = Z_1) \\
&+ P(X_1 = \alpha, Z_1 = -\beta)P(X_2 = X_1)P(Z_2 \neq Z_1) \\
&+ P(X_1 = -\alpha, Z_1 = \beta)P(X_2 \neq X_1)P(Z_2 = Z_1) \\
&+ P(X_1 = -\alpha, Z_1 = -\beta)P(X_2 \neq X_1)P(Z_2 \neq Z_1) \\
=&\frac{1}{4}(1 - P(X_2 \neq X_1))[P(Z_2 = Z_1) + P(Z_2 \neq Z_1)] \\
&+ \frac{1}{4}P(X_2 \neq X_1)[P(Z_2 = Z_1) + P(Z_2 \neq Z_1)] = \frac{1}{4}.
\end{aligned}$$

The same applies recursively for the other joint probabilities so that it applies to all n. Therefore, if one were to calculate the covariance or Mutual Information values between these processes they will all be the same. In fact both of the measures would be 0. Since $E(XY) = E(X)E(Y)$ due to independent probabilities, the covariance is,

$$\Gamma(X, Y) = E(XY) - E(X)E(Y) = 0.$$

The Mutual Information will be

$$I(X, Y) = E\left[log\,\frac{P(X, Y)}{P(Y)P(X)}\right] = E\left[log\,\frac{\frac{1}{4}}{\frac{1}{2}\frac{1}{2}}\right] = 0.$$

Evidently, in this case, covariance and Mutual Information are unable to provide any information regarding the relationship between the processes.

In relation to probabilities of the amended Ising model outlined in subsection (6.1.3), the transition probabilities of processes can be written as

$$P(X_n = \alpha | X_{n-1} = \beta) = \begin{cases} 1 - \mu_X P(Z_{n-1} = 1) = 1 - \frac{1}{2}\mu_X, & \text{if } \alpha = \beta \\ \mu_X P(Z_{n-1} = 1) = \frac{1}{2}\mu_X, & \text{if } \alpha \neq \beta \end{cases}$$

$$P(Y_n = \alpha | Y_{n-1} = \beta) = \begin{cases} 1 - \mu_Y P(Z_{n-1} = 1) = 1 - \frac{1}{2}\mu_Y, & \text{if } \alpha = \beta \\ \mu_Y P(Z_{n-1} = 1) = \frac{1}{2}\mu_Y, & \text{if } \alpha \neq \beta \end{cases}$$

and

$$P(Z_n = \alpha | Z_{n-1} = \beta) = \begin{cases} 1 - \mu_Z, & \text{if } \alpha = \beta \\ \mu_Z, & \text{if } \alpha \neq \beta \end{cases}$$

where $P(Z_{n-1} = 1) = \frac{1}{2}$ as per discussion above. If a certain process does not affect the other process at time $n-1$, for example Y_{n-1} does not influence X_n, then we have that

$$P(X_n = \alpha | X_{n-1} = \beta, Y_{n-1} = \gamma) = P(X_n = \alpha | X_{n-1} = \beta)$$

for any $\alpha, \beta, \gamma \in A$. This is true for all the other possibilities except when we condition on Z, i.e $P(X_n = \alpha | X_{n-1} = \beta, Z_{n-1} = \gamma)$ and $P(Y_n = \alpha | Y_{n-1} = \beta, Z_{n-1} = \gamma)$. Recall that this is because we have imposed that X_{n-1} and Y_{n-1} can only change if $Z_{n-1} = 1$.

Let $Q^{(\tau)}_{sgn(\gamma)} = P(\text{ condition fulfilled } | Z_{n-\tau} = \gamma) = P(Z_{n-1} = 1 | Z_{n-1} = \gamma)$ as in equation (6.2). Given that $Z_{n-1} = -1$, we get $Q^{(1)}_{-} = P(Z_{n-1} = 1 | Z_{n-1} = -1) = 0$ so that

$$P(X_n = \alpha | X_{n-1} = \beta, Z_{n-1} = -1) = \begin{cases} 1 - \mu_X Q^{(1)}_{-} = 1, & \text{if } \alpha = \beta \\ \mu_X Q^{(1)}_{-} = 0, & \text{if } \alpha \neq \beta \end{cases}$$

and

$$P(Y_n = \alpha | Y_{n-1} = \beta, Z_{n-1} = -1) = \begin{cases} 1 - \mu_Y Q^{(1)}_{-} = 1, & \text{if } \alpha = \beta \\ \mu_Y Q^{(1)}_{-} = 0, & \text{if } \alpha \neq \beta. \end{cases}$$

Otherwise if $Z_{n-1} = 1$ we get $Q^{(1)}_{+} = P(Z_{n-1} = 1 | Z_{n-1} = 1) = 1$ so that

$$P(X_n = \alpha | X_{n-1} = \beta, Z_{n-1} = 1) = \begin{cases} 1 - \mu_X Q^{(1)}_{+} = 1 - \mu_X, & \text{if } \alpha = \beta \\ \mu_X Q^{(1)}_{+} = \mu_X, & \text{if } \alpha \neq \beta \end{cases}$$

and

$$P(Y_n = \alpha | Y_{n-1} = \beta, Z_{n-1} = 1) = \begin{cases} 1 - \mu_Y Q^{(1)}_{+} = 1 - \mu_Y, & \text{if } \alpha = \beta \\ \mu_Y Q^{(1)}_{+} = \mu_Y, & \text{if } \alpha \neq \beta. \end{cases}$$

Therefore if we divide these probabilities with the corresponding transition probabilities,

we get that,

$$\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-1}=-1)}{P(X_n=\alpha|X_{n-1}=\beta)}=\begin{cases}\frac{1}{1-\frac{1}{2}\mu_X}=\frac{2}{2-\mu_X}, & \text{if } \alpha=\beta\\ 0, & \text{if } \alpha\neq\beta,\end{cases}$$

$$\frac{P(Y_n=\alpha|Y_{n-1}=\beta,Z_{n-1}=-1)}{P(Y_n=\alpha|Y_{n-1}=\beta)}=\begin{cases}\frac{1}{1-\frac{1}{2}\mu_Y}=\frac{2}{2-\mu_Y}, & \text{if } \alpha=\beta\\ 0, & \text{if } \alpha\neq\beta,\end{cases}$$

$$\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-1}=1)}{P(X_n=\alpha|X_{n-1}=\beta)}=\begin{cases}\frac{1-\mu_X}{1-\frac{1}{2}\mu_X}=\frac{2(1-\mu_X)}{2-\mu_X}, & \text{if } \alpha=\beta\\ \frac{\mu_X}{\frac{1}{2}\mu_X}=2, & \text{if } \alpha\neq\beta,\end{cases}$$

and

$$\frac{P(Y_n=\alpha|Y_{n-1}=\beta,Z_{n-1}=1)}{P(Y_n=\alpha|Y_{n-1}=\beta)}=\begin{cases}\frac{1-\mu_Y}{1-\frac{1}{2}\mu_Y}=\frac{2(1-\mu_Y)}{2-\mu_Y}, & \text{if } \alpha=\beta\\ \frac{\mu_Y}{\frac{1}{2}\mu_Y}=2, & \text{if } \alpha\neq\beta.\end{cases}$$

The Transfer Entropy is defined to quantify this ratio.

7.1.2 Transfer Entropy on the simple model

Recall from equation (4.5) that $T_{ZX}^{(t_Z)}=E\left[\,log\,\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-t_Z}=\gamma)}{P(X_n=\alpha|X_{n-1}=\beta)}\right]$. Therefore the Transfer Entropy $T_{ZX}^{(1)}$ is calculated as,

$$\begin{aligned}T_{ZX}^{(1)}&=E\left[\,log\,\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-1}=\gamma)}{P(X_n=\alpha|X_{n-1}=\beta)}\right]\\&=\sum_{\alpha\in A}\sum_{\beta\in A}\sum_{\gamma\in A}P(X_n=\alpha,X_{n-1}=\beta,Z_{n-1}=\gamma)\,log\,\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-1}=\gamma)}{P(X_n=\alpha|X_{n-1}=\beta)}\\&=2\left[\frac{1}{4}(1)\,log\,\frac{2}{2-\mu_X}+0+\frac{1}{4}(1-\mu_X)\,log\,\frac{2(1-\mu_X)}{2-\mu_X}+\frac{1}{4}(\mu_X)\,log\,2\right]\\&=log\,2+\frac{1}{2}\left[(1-\mu_X)\,log\,(1-\mu_X)-(2-\mu_X)\,log\,(2-\mu_X)\right]\end{aligned}\tag{7.1}$$

where

$$P(X_n=\alpha,Z_{n-1}=\gamma,X_{n-1}=\beta)=\frac{1}{4}P(X_n=\alpha|Z_{n-1}=\gamma,X_{n-1}=\beta).$$

In the same way, we get that

$$T_{ZY}^{(1)} = log\, 2 + \frac{1}{2}\left[(1-\mu_Y)\, log\,(1-\mu_Y) - (2-\mu_Y)\, log\,(2-\mu_Y)\right].$$

From Figure (7.3) where equation (7.1) is analytically plotted, we can say that $T_{ZX}^{(1)} \neq 0$

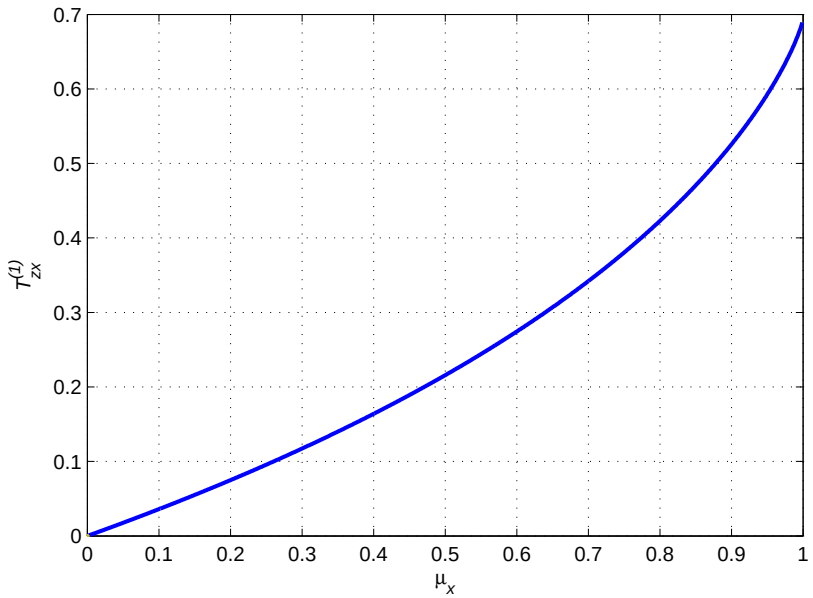

Figure 7.3: Analytical values of $T_{ZX}^{(1)}$ versus μ_X in equation (7.1) of the simple model with $t_Z = 1$. $T_{ZX}^{(1)}$ in equation (7.1) is a monotonically increasing function of μ_X.

except when $\mu_X = 0$ in which case, X becomes a constant. If we assign a value for example $\mu_X = \frac{1}{2}$, we get that $T_{ZX}^{(1)} = T_{ZY}^{(1)} = \frac{3}{2}\, log\, 2 - \frac{3}{4}\, log\, 3 \approx 0.2158$ whereas $T_{XZ}^{(1)} = T_{YZ}^{(1)} = 0$ which clearly indicates a causal direction from $Z \rightarrow X$ and $Z \rightarrow Y$. Analytically, the values of $T_{XY}^{(1)} = T_{YX}^{(1)} = 0$ since the ratios will be 1. This correctly indicates that X and Y does not have a causal relationship at $t_Z = 1$.

7.2 The general model

Previously in the simple model, we only had two states in the model since $A = \{-1, 1\}$, but in the real world, we do not always have this luxury. Let $n_s \geq 2$ be the number of states we have in the model, and define $A = \{1, ..., n_s\}$ as the set of possible states. Note that

the simplest case $n_s = 2$ is equivalent to the previous simple model. As before, let μ_X, μ_Y and μ_Z, be the independent probabilities for the stochastic swaps of the variables X, Y and Z at every time step respectively. Again we impose a special restriction on X and Y such that they are only allowed to do the stochastic swap with probability μ_X and μ_Y if the state of Z_{n-t_Z} fulfills a certain condition. This restriction means that X and Y can only change states if Z is in the conditioned state at time step $n - t_Z$ thus creating a 'dependence' on Z.

The processes are initialized randomly and independently so that the probabilities are uniform i.e. $P(X_1 = 1) = \cdots = P(Z_1 = n_s) = \frac{1}{n_s}$. We set the model to be such that if a process chooses to change it must choose one of the other states equally, thus we have that $P(X_2 = \alpha | X_1 = \beta, \alpha \neq \beta) = \frac{1}{n_s - 1} P(X_2 \neq X_1)$, since $\frac{1}{n_s-1}$ is the probability that X chooses α given that it must change. Therefore for $\alpha, \beta \in A$,

$$\begin{aligned} P(X_2 = \alpha) &= P(X_1 = \alpha)P(X_2 = X_1) + P(X_1 \neq \alpha)P(X_2 = \alpha | X_1 = \beta, \alpha \neq \beta) \\ &= \frac{1}{n_s}(1 - P(X_2 \neq X_1)) + \frac{n_s - 1}{n_s}\frac{1}{n_s - 1}P(X_2 \neq X_1) \\ &= \frac{1}{n_s}(1 - P(X_2 \neq X_1) + P(X_2 \neq X_1)) = \frac{1}{n_s}. \end{aligned}$$

Applying this recursively gives $P(X_n = 1) = ... = P(X_n = n_s) = \frac{1}{n_s}$ for any n. The same goes for Y and Z. It also follows that since the processes are initialized randomly and independently, for $n = 1$ all the joint probabilities are the product of the marginal probabilities thus the value of the joint probability of two processes becomes $\frac{1}{n_s^2}$. The joint probability of three processes become $\frac{1}{n_s^3}$. This generalizes to any number of processes, so that the joint probability becomes simply a product of the marginal ones. For example $P(X_1 = \alpha, Z_1 = \beta) = P(X_1 = \alpha)P(Z_1 = \beta) = \frac{1}{n_s^2}$ for any $\alpha, \beta \in A$. We make use of the equation

$$P(X_1 = \alpha, Z_1 \neq \beta) = \sum_{\gamma \neq \beta} P(X_1 = \alpha, Z_1 = \gamma) = \sum_{\gamma \neq \beta} \frac{1}{n_s^2} = \frac{n_s - 1}{n_s^2}$$

to get

$$\begin{aligned}P(X_2 =&\alpha, Z_2 = \beta) = P(X_1 = \alpha, Z_1 = \beta)P(X_2 = X_1)P(Z_2 = Z_1)\\ &+ P(X_1 = \alpha, Z_1 \neq \beta)P(X_2 = X_1)\frac{1}{n_s - 1}P(Z_2 \neq Z_1)\\ &+ P(X_1 \neq \alpha, Z_1 = \beta)\frac{1}{n_s - 1}P(X_2 \neq X_1)P(Z_2 = Z_1)\\ &+ P(X_1 \neq \alpha, Z_1 \neq \beta)\frac{1}{n_s - 1}P(X_2 \neq X_1)\frac{1}{n_s - 1}P(Z_2 \neq Z_1) = \frac{1}{n_s^2}\end{aligned}$$

for $\alpha, \beta \in A$. Recursively we get that it applies to all n, and for joint probabilities of other variables as well, therefore joint probabilities are all $\frac{1}{n_s^2}$ regardless of the values of the transition probabilities. Also, if we were to calculate the covariance, Mutual Information or even conditional Mutual Information values between these processes they will all be 0.

To get nonzero values one has to look at the transition probabilities,

$$P(X_n = \alpha|X_{n-1} = \beta) = \begin{cases}1 - \mu_X\Omega & \text{if } \alpha = \beta\\ \frac{1}{n_s-1}\mu_X\Omega & \text{if } \alpha \neq \beta\end{cases}$$

$$P(Y_n = \alpha|Y_{n-1} = \beta) = \begin{cases}1 - \mu_Y\Omega & \text{if } \alpha = \beta\\ \frac{1}{n_s-1}\mu_Y\Omega & \text{if } \alpha \neq \beta.\end{cases}$$

and

$$P(Z_n = \alpha|Z_{n-1} = \beta) = \begin{cases}1 - \mu_Z & \text{if } \alpha = \beta\\ \frac{1}{n_s-1}\mu_Z & \text{if } \alpha \neq \beta\end{cases}$$

where $\Omega = P($ condition fulfilled $)$ such that

$$P(X_n \neq X_{n-1}) = \sum_{\beta\neq\alpha} P(X_n = \alpha|X_{n-1} = \beta) = \mu_X\Omega$$

and similarly $P(Y_n \neq Y_{n-1}) = \mu_Y\Omega$. We can change to what extend the 'dependence' on Z is by altering Ω. To understand how the values of μ_Z affects the value of $T_{ZX}^{(\tau)}$ through $Q_{sgn(\gamma)}^{(\tau)}$, we will need to look at the relationship between Ω and Q.

7.2.1 The relationship between Ω and Q

Recall that for the simple model and the amended Ising model where there are only two possible states, we have defined $Q^{(\tau)}_{sgn(\gamma)} = P(\text{ condition fulfilled } | Z_{n-1} = \gamma) = P(Z_{n-t_Z} = 1 | Z_{n-\tau} = \gamma)$ in equation (6.2) where

$$sgn(\gamma) = \begin{cases} + & \text{if } \gamma = 1 \\ - & \text{if } \gamma = -1. \end{cases} \tag{7.2}$$

Now in the general model where $\gamma \in A = \{1, \ldots, n_s\}$ are all positive integers, the possible states are different. The value of $Q^{(\tau)}_{sgn(\gamma)}$ will depend on γ, and in our model here, particularly whether or not $Z_{n-t_z} = \gamma$ satisfies the condition. One can divide the possible states γ of all the processes into two groups such that

$$G_U = \{\gamma \in A, Z_{n-t_Z} = \gamma \text{ fulfills the condition}\} \quad \text{and}$$

$$G_D = \{\gamma \in A, Z_{n-t_Z} = \gamma \text{ does not fulfill the condition}\}.$$

Note that $|G_U| = n_s\Omega$ and $|G_D| = n_s(1-\Omega)$ since $\Omega = P(\text{ condition fulfilled })$ such that Ω can be interpreted as the proportion of states of Z that fulfill the condition. Q represents the probability that the condition is fulfilled given current knowledge at time τ such that $Q^{(\tau)}_{sgn(\gamma)} = P(\text{ condition fulfilled } | \text{ knowledge at time } \tau)$. Due to equiprobability of spins and uniform initial distribution, for any τ there are only two possible values of $Q^{(\tau)}_{sgn(\gamma)}$, one for $\gamma \in G_U$ and one for $\gamma \in G_D$. Therefore we need to redefine $sgn(\gamma)$ such that

$$sgn(\gamma) = \begin{cases} + & \text{if } \gamma \in G_U \\ - & \text{if } \gamma \in G_D \end{cases} \tag{7.3}$$

to get

$$Q^{(\tau)}_{sgn(\gamma)} = \begin{cases} Q^{(\tau)}_{+} & \text{if } \gamma \in G_U \\ Q^{(\tau)}_{-} & \text{if } \gamma \in G_D. \end{cases} \tag{7.4}$$

For the general model, we shall define $Q^{(\tau)}_{sgn(\gamma)} = P(\text{ condition fulfilled } | Z_{n-\tau} = \gamma)$ with the $sgn(\gamma)$ as in equation (7.3).

The relationship between $Q^{(\tau)}_{sgn(\gamma)}$ and Ω can be defined using the formula for total probability $P(B) = \sum_\gamma P(B|Z = \gamma)P(Z = \gamma)$. Let $B = \{$ condition fulfilled $\}$ and using the fact that $P(Z_{n-\tau} = \gamma) = \frac{1}{n_s}$, we get that

$$\Omega = P(B) = \sum_\gamma P(B|Z_{n-\tau} = \gamma)P(Z_{n-\tau} = \gamma) = \frac{1}{n_s}\sum_\gamma Q^{(\tau)}_{sgn(\gamma)}. \tag{7.5}$$

Due to the sole dependence of Z on μ_Z, $\mu_Z = \frac{n_s-1}{n_s}$ will make the transition probability of Z uniform such that $P(Z_n = \alpha|Z_{n-1} = \beta) = \frac{1}{n_s}$ for any n since we have that

$$P(Z_n = \alpha|Z_{n-1} = \beta) = \begin{cases} 1 - \mu_Z = 1 - \frac{n_s-1}{n_s} = \frac{1}{n_s} & \text{if } \alpha = \beta \\ \frac{1}{n_s-1}\mu_Z = \frac{1}{n_s-1}\frac{n_s-1}{n_s} = \frac{1}{n_s} & \text{if } \alpha \neq \beta \end{cases}$$

for any $\alpha, \beta \in A = \{1, \cdots, n_s\}$. Consequently, $\mu_Z = \frac{n_s-1}{n_s}$ also makes all values of $Q^{(\tau)}_{sgn(\gamma)}$ uniform so that equation (7.5) becomes

$$\Omega = \frac{1}{n_s}\sum_\gamma Q^{(\tau)}_{sgn(\gamma)} = \frac{1}{n_s} n_s Q^{(\tau)}_{sgn(\gamma)} = Q^{(\tau)}_{sgn(\gamma)}. \tag{7.6}$$

Therefore on the model when the $\mu_Z = \frac{n_s-1}{n_s}$, we have that $\Omega = Q^{(\tau)}_{sgn(\gamma)}$ for any $\tau = t_Z$.

For any μ_Z, the relationship between $Q^{(\tau)}_+$ and $Q^{(\tau)}_-$ can be derived from equation (7.5) where

$$n_s\Omega = \sum_\gamma Q^{(\tau)}_{sgn(\gamma)} = \sum_{\gamma \in G_U} Q^{(\tau)}_{sgn(\gamma)} + \sum_{\gamma \in G_D} Q^{(\tau)}_{sgn(\gamma)} = |G_U|Q^{(\tau)}_+ + |G_D|Q^{(\tau)}_- \tag{7.7}$$

$$n_s\Omega = n_s\Omega\ Q^{(\tau)}_+ + n_s(1-\Omega)Q^{(\tau)}_-$$

$$\Omega(1 - Q^{(\tau)}_+) = (1-\Omega)Q^{(\tau)}_-$$

Note that when $n_s = 2$ (hence $\Omega = \frac{1}{2}$) this simplifies to $Q^{(\tau)}_+ + Q^{(\tau)}_- = 1$.

7.2.2 Transfer Entropy on the model

Using $Q^{(\tau)}_{sgn(\gamma)}$ as in equation (7.4) we have that

$$\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-\tau}=\gamma)}{P(X_n=\alpha|X_{n-\tau}=\beta)}=\begin{cases}\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{1-\mu_X\Omega} & \text{if } \alpha=\beta\\ \frac{\frac{1}{n_s-1}\mu_X Q^{(\tau)}_{sgn(\gamma)}}{\frac{1}{n_s-1}\mu_X\Omega}=\frac{Q^{(\tau)}_{sgn(\gamma)}}{\Omega} & \text{if } \alpha\neq\beta,\end{cases}$$

which gives us

$$\begin{aligned}
T^{(\tau)}_{ZX} =& \sum_\alpha\sum_\beta\sum_\gamma P(X_n=\alpha,X_{n-1}=\beta,Z_{n-\tau}=\gamma)log\,\frac{P(X_n=\alpha|X_{n-1}=\beta,Z_{n-\tau}=\gamma)}{P(X_n=\alpha|X_{n-1}=\beta)}\\
=&|\{X_n=X_{n-1}\}|\sum_\gamma\left[\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{n_s^2}\,log\,\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{1-\mu_X\Omega}\right]\\
&+|\{X_n\neq X_{n-1}\}|\sum_\gamma\left[\frac{\frac{1}{n_s-1}\mu_X Q^{(\tau)}_{sgn(\gamma)}}{n_s^2}\,log\,\frac{Q^{(\tau)}_{sgn(\gamma)}}{\Omega}\right]\\
=&n_s\sum_\gamma\left[\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{n_s^2}\,log\,\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{1-\mu_X\Omega}\right]\\
&+n_s(n_s-1)\sum_\gamma\left[\frac{\frac{1}{n_s-1}\mu_X Q^{(\tau)}_{sgn(\gamma)}}{n_s^2}\,log\,\frac{Q^{(\tau)}_{sgn(\gamma)}}{\Omega}\right]\\
=&\frac{1}{n_s}\sum_{\gamma\in G_U}\left[(1-\mu_X Q^{(\tau)}_{sgn(\gamma)})\,log\,\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{sgn(\gamma)}\,log\,\frac{Q^{(\tau)}_{sgn(\gamma)}}{\Omega}\right]\\
&+\frac{1}{n_s}\sum_{\gamma\in G_D}\left[(1-\mu_X Q^{(\tau)}_{sgn(\gamma)})\,log\,\frac{1-\mu_X Q^{(\tau)}_{sgn(\gamma)}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{sgn(\gamma)}\,log\,\frac{Q^{(\tau)}_{sgn(\gamma)}}{\Omega}\right]\\
=&\frac{1}{n_s}(n_s\Omega)\left[(1-\mu_X Q^{(\tau)}_{+})\,log\,\frac{1-\mu_X Q^{(\tau)}_{+}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{+}\,log\,\frac{Q^{(\tau)}_{+}}{\Omega}\right]\\
&+\frac{1}{n_s}n_s(1-\Omega)\left[(1-\mu_X Q^{(\tau)}_{-})\,log\,\frac{1-\mu_X Q^{(\tau)}_{-}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{-}\,log\,\frac{Q^{(\tau)}_{-}}{\Omega}\right]\\
=&\Omega\left[(1-\mu_X Q^{(\tau)}_{+})\,log\,\frac{1-\mu_X Q^{(\tau)}_{+}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{+}\,log\,\frac{Q^{(\tau)}_{+}}{\Omega}\right]\\
&+(1-\Omega)\left[(1-\mu_X Q^{(\tau)}_{-})\,log\,\frac{1-\mu_X Q^{(\tau)}_{-}}{1-\mu_X\Omega}+\mu_X Q^{(\tau)}_{-}\,log\,\frac{Q^{(\tau)}_{-}}{\Omega}\right]
\end{aligned}\tag{7.8}$$

where we used the Bayes theorem i.e

$$P(X_n = \alpha, Z_{n-1} = \gamma, X_{n-1} = \beta) = \frac{1}{n_s^2} P(X_n = \alpha | Z_{n-1} = \gamma, X_{n-1} = \beta).$$

Due to independence, if Y were to be conditioned on X we would have that

$$\frac{P(Y_n = \alpha | Y_{n-1} = \beta, X_{n-\tau} = \gamma)}{P(Y_n = \alpha | Y_{n-1} = \beta)} = \frac{P(Y_n = \alpha | Y_{n-1} = \beta)}{P(Y_n = \alpha | Y_{n-1} = \beta)} = 1.$$

Therefore for values other than when X and Y conditioned on Z, this ratio will yield 1. This renders $T_{XZ}^{(\tau)} = T_{YZ}^{(\tau)} = T_{YX}^{(\tau)} = T_{XY}^{(\tau)} = 0$. And if we get that $T_{ZX}^{(\tau)} = T_{ZX}^{(\tau)} \neq 0$, we can say that Transfer Entropy indicates 'causality' from Z to X and Z to Y, which is exactly what we want. In a similar manner for $\alpha, \beta, \gamma \in A$ we have that

$$\frac{P(Y_n = \alpha | Y_{n-1} = \beta, Z_{n-\tau} = \gamma)}{P(Y_n = \alpha | Y_{n-1} = \beta)} = \begin{cases} \frac{1-\mu_Y Q_{sgn(\gamma)}^{(\tau)}}{1-\mu_Y \Omega} & \text{if } \alpha = \beta \\ \frac{\frac{1}{n_s-1}\mu_Y Q_{sgn(\gamma)}^{(\tau)}}{\frac{1}{n_s-1}\mu_Y \Omega} = \frac{Q_{sgn(\gamma)}^{(\tau)}}{\Omega} & \text{if } \alpha \neq \beta \end{cases}$$

such that $T_{ZY}^{(\tau)}$ in exactly like equation (7.8) except that μ_X is replaced with μ_Y.

When $\tau = t_Z$ we have that $Q_{sgn(\gamma)}^{(t_Z)}$ is either 0 or 1 since the condition was placed at $n - t_Z$. More specifically we will have that $Q_+^{(t_Z)} = 1$ and that $Q_-^{(t_Z)} = 0$. Putting these two values in equation (7.8) we obtain

$$\begin{aligned} T_{ZX}^{(t_Z)} =& \Omega \left[(1 - \mu_X Q_+^{(t_Z)}) \, log \, \frac{1 - \mu_X Q_+^{(t_Z)}}{1 - \mu_X \Omega} + \mu_X Q_+^{(t_Z)} \, log \, \frac{Q_+^{(t_Z)}}{\Omega} \right] \\ &+ (1 - \Omega) \left[(1 - \mu_X Q_-^{(t_Z)}) \, log \, \frac{1 - \mu_X Q_-^{(t_Z)}}{1 - \mu_X \Omega} + \mu_X Q_-^{(t_Z)} \, log \, \frac{Q_-^{(t_Z)}}{\Omega} \right] \\ =& \; \Omega(1 - \mu_X) \, log \, \frac{1 - \mu_X}{1 - \mu_X \Omega} + \Omega \mu_X \, log \, \frac{1}{\Omega} + (1 - \Omega) \, log \, \frac{1}{1 - \mu_X \Omega}. \end{aligned} \tag{7.9}$$

When $\Omega = 0.5$ and $t_Z = 1$ in equation (7.9), we have that the formula is exactly like the equation (7.1) for the simple model. This is shown in Figure (7.4) where it is the red dotted line (more on this special case later) is exactly Figure (7.3). $T_{ZX}^{(t_Z)}$ values for $\Omega = 0.25$ and $\Omega = 0.75$ in Figure (7.4) converges since equation (7.9) becomes equal for any pair of Ω and $1 - \Omega$ at $\mu_X = 1$. In other words equation (7.9) is symmetrical over Ω when $\mu_X = 1$.

Also notice that for $\Omega = 1$ we have that $T_{ZX}^{(t_Z)} = 0$ since this means that the condition is fulfilled all the time which is equal to Z being not restrictive at all.

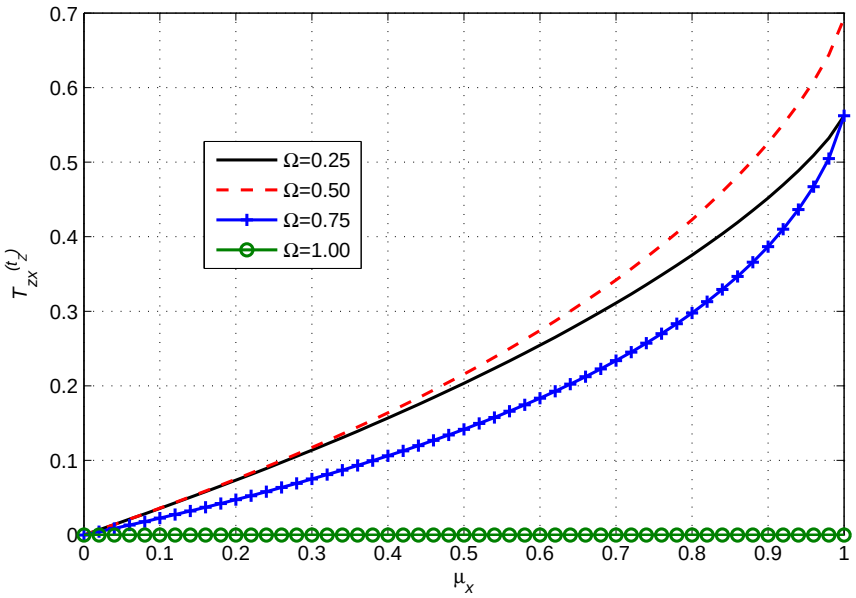

Figure 7.4: Analytical value of $T_{ZX}^{(t_Z)}$ versus μ_X in equation (7.9) of the general model for different values of $\Omega = P($ condition fulfilled $)$. For fixed Ω, $T_{ZX}^{(t_Z)}$ in equation (7.9) is a monotonically increasing function of μ_X .

7.2.3 Transfer Entropy for causal lag detection

If there exists a specific causal lag in the model where either X, Y or Z causes each other, then the Transfer Entropy should be able to detect the direction and the exact causal lag using equation (7.8). In this model where we have imposed that Z 'causes' X and Y at causal lag t_Z, we shall show that $T_{ZX}^{(\tau)} \leq T_{ZX}^{(t_Z)}$ (and similarly $T_{ZY}^{(\tau)} \leq T_{ZY}^{(t_Z)}$). Consequently the largest value that we get for $T_{ZX}^{(\tau)}$ (and $T_{ZY}^{(\tau)}$) indicates the actual causal lag. For the most part of the thesis, we will mostly focus on $T_{ZX}^{(\tau)}$, fully realizing the fact the $T_{ZY}^{(\tau)}$ has exactly the same probability as $T_{ZX}^{(\tau)}$ and therefore what applies to the relationship between X and Z applies equally to the relationship between Y and Z.

We have seen that $\mu_Z = \frac{n_s - 1}{n_s}$ (resulting in uniform transition probability) leads to $Q_{sgn(\gamma)}^{(\tau)} = \Omega$ in equation (7.6), which in turn renders $T_{ZX}^{(\tau)} = 0$ whenever $\tau \neq t_Z$. One

can clearly see this by substituting $Q^{(\tau)}_{sgn(\gamma)} = \Omega$ in equation (7.8). This is due to the fact that $Q^{(\tau)}_{sgn(\gamma)} = P(\text{ condition fulfilled } | Z_{n-\tau} = \gamma)$ and the condition is manifested only at Z_{n-t_Z} so that when $\tau = t_Z$, $Q^{(t_Z)}_{sgn(\gamma)}$ is either 1 or 0 hence resulting in equation (7.9). We can clearly use the fact that $T^{(\tau)}_{ZX} \neq 0$ only at $\tau = t_Z$ to detect time lags since this obviously implies $T^{(\tau)}_{ZX} \leq T^{(t_Z)}_{ZX}$. To illustrate, let $n_s = 2$ so that $A = \{1, 2\}$. Note that the probabilities are equivalent to the simple model. Let the condition be $Z_{n-t_Z} = 1$ so that $\Omega = P(Z_{n-t_Z} = 1) = \frac{1}{2}$ and $Q^{(\tau)}_{sgn(\gamma)} = P(Z_{n-t_Z} = 1 | Z_{n-\tau} = \gamma)$. Thus (7.8) becomes

$$T^{(\tau)}_{ZX} = \frac{1}{2}\left[(1-\mu_X Q^{(\tau)}_{+})\, log\, \frac{2(1-\mu_X Q^{(\tau)}_{+})}{2-\mu_X} + \mu_X Q^{(\tau)}_{+}\, log\, 2Q^{(\tau)}_{+}\right] + \frac{1}{2}\left[(1-\mu_X Q^{(\tau)}_{-})\, log\, \frac{2(1-\mu_X Q^{(\tau)}_{-})}{2-\mu_X} + \mu_X Q^{(\tau)}_{-}\, log\, 2Q^{(\tau)}_{-}\right]. \tag{7.10}$$

When $\mu_Z = \frac{n_s - 1}{n_s} = \frac{1}{2}$, transition probability is uniformly distributed thus we get Figure

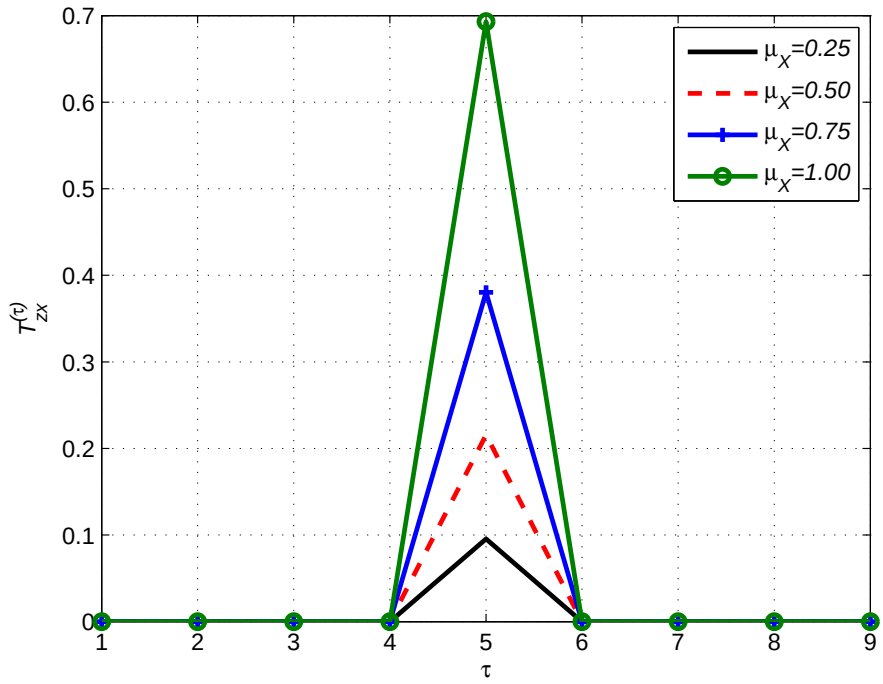

Figure 7.5: Analytical $T^{(\tau)}_{ZX}$ versus τ in equation (7.10) with fixed values of $\mu_Z = \frac{1}{2}$, $t_Z = 5$ and $n_s = 2$ (so that $\Omega = \frac{1}{2}$). μ_X effects the values of $T^{(\tau)}_{ZX}$ at $\tau = t_Z$.

(7.5) which is equation (7.10) plotted for $\tau = 1, \cdots, 9$ with four different μ_X values. $t_Z = 5$ was chosen for illustration purposes. The values of $T_{ZX}^{(\tau)}$ in this figure are only dependent μ_X and time lag detection is straight forward. We get that $T_{ZX}^{(t_Z)}$ increases as μ_X increases in Figure (7.5). Setting $\mu_Z = \frac{n_s-1}{n_s}$ causes all $Q_{sgn(\gamma)}^{(\tau)} = \Omega$ so that all $T_{ZX}^{(\tau)} = 0$ for any $\tau \neq t_Z$. This can be verified by a brief inspection of equation (7.8).

However for varying μ_Z values we get quite a different picture. Only when $\mu_Z \neq \frac{n_s-1}{n_s}$, do we get cases where $T_{ZX}^{(\tau)} \neq 0$ when $\tau \neq t_Z$. Figure (7.6) is equation (7.10) plotted for $\tau = 1, \cdots, 9$ with four different μ_Z values when $\mu_X = \frac{1}{2}$ is fixed. Fixing $\mu_X = \frac{n_s-1}{n_s} = \frac{1}{2}$ makes the transition probability of X uniform (except for the bit influenced by Z). The red dotted line (when both $\mu_X = \frac{1}{2}$ and $\mu_Z = \frac{1}{2}$) in Figure (7.5) and Figure (7.6) are equivalent. Again $t_Z = 5$ was chosen for illustration purposes. The fact that $T_{ZX}^{(t_Z)}$ values only depend on μ_X is manifested in Figure (7.6) where μ_X is fixed and we can see that the peaks converge at a single value. We can also clearly see in Figure (7.6) that when $\mu_Z = 1$

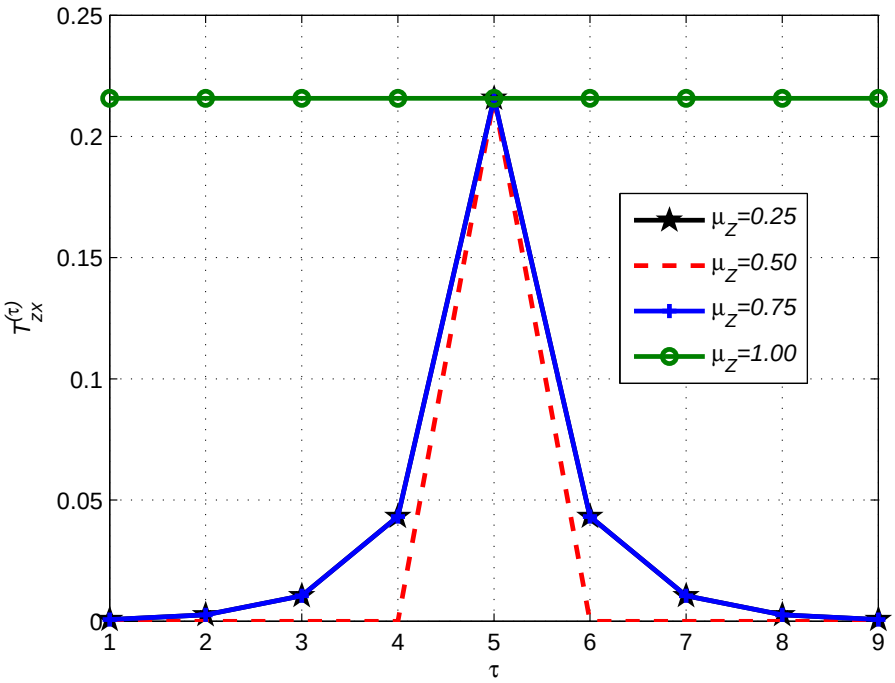

Figure 7.6: Analytical $T_{ZX}^{(\tau)}$ versus τ in equation (7.10) with fixed values of $\mu_X = \frac{1}{2}$, $t_Z = 5$ and $n_s = 2$ (so that $\Omega = \frac{1}{2}$). μ_Z only effects the values of $T_{ZX}^{(\tau)}$ at $\tau \neq t_Z$.

for any τ, $T_{ZX}^{(\tau)} = T_{ZX}^{(t_Z)}$. This situation is unique for $n_s = 2$ (the simple model) since $\mu_Z = 1$ leads to Z being deterministic (changes at every time step). This only happens when there are only two possible states due to the fact that if it needs to change, it only has one other spin to go to (for $n_s > 2$, more states are available to choose from thus retaining the stochastic element). Therefore given any single value of Z at any time step, one will be able to determine the value of Z_{n-t_Z}. For example if $Z_1 = 1$, one would know that $Z_2 = 2$, $Z_3 = 1$ and so on. Consequently for $\mu_Z = 1$, for $\alpha, \beta, \gamma \in A = \{-1, 1\}$ we have that

$$\begin{aligned} T_{ZX}^{(\tau)} &= E\left[log\, \frac{P(X_n = \alpha | X_{n-1} = \beta, Z_{n-\tau} = \gamma)}{P(X_n = \alpha | X_{n-1} = \beta)} \right] \\ &= E\left[log\, \frac{P(X_n = \alpha | X_{n-1} = \beta, Z_{n-t_Z} = \gamma)}{P(X_n = \alpha | X_{n-1} = \beta)} \right] = T_{ZX}^{(t_Z)}. \end{aligned}$$

Which leads us to conclude that if the variable we are conditioning on (in this case Z) is deterministic, the Transfer Entropy value is independent of the time lags in the simple model. This is an interesting case in point, since $\mu_Z = 1$ makes Z deterministic and having $\mu_X = \frac{1}{2}$ keeps X stochastic while still depending on Z, thus this is a case where a stochastic process is dependent on a deterministic process. Transfer Entropy gives a clear direction from $Z \to X$ since $T_{XZ}^{(\tau)} = 0$ and $T_{ZX}^{(\tau)} \neq 0$ for any τ. Therefore, although causal lag detection cannot be established, in this case Transfer Entropy does indeed succeed in giving a direction despite one of the processes being deterministic. The original definition in [89] excluded cases when one or both process is deterministic as previously discussed in subsection (4.3.1).

Essentially Figure (7.6) depicts how the distance from t_Z influences the value of $T_{ZX}^{(\tau)}$. Therefore let $\nu = |\tau - t_Z|$ be the distance of τ from t_Z. When $\nu = 0$ (for example $\tau = 5$ in Figure (7.6)) since $Q_+^{(\tau)} = P(Z_{n-t_Z} = 1 | Z_{n-\tau} = 1) = 1$ and $Q_-^{(\tau)} = 0$, equation (7.10) will simply become

$$\begin{aligned} T_{ZX}^{(t_Z)} =& \frac{1}{2}\left[(1-\mu_X)\, log\, \frac{2(1-\mu_X)}{2-\mu_X} + \mu_X\, log\, 2 + log\, \frac{2}{2-\mu_X} \right] \\ =& log\, 2 + \frac{1}{2}\left[(1-\mu_X)\, log\, (1-\mu_X) - (2-\mu_X)\, log\, (2-\mu_X) \right], \end{aligned}$$

which coincides with equation (7.1) of the simple model. This equation is independent of

μ_Z. However if $\nu = 1$, for example $\tau = 4, 6$ in Figure (7.6), the values of Q are now different since τ is one time step away from t_Z so that

$$Q_+^{(\tau)} = P(Z_{n-t_Z} = 1 | Z_{n-\tau} = 1) = 1 - \mu_Z \text{ and } Q_-^{(\tau)} = 1 - Q_+^{(\tau)} = \mu_Z.$$

Therefore now equation (7.10) is dependent on μ_Z since

$$\begin{aligned} T_{ZX}^{(\tau)} =& \frac{1}{2}\left[(1-\mu_X Q_+^{(\tau)})\, log\, \frac{2(1-\mu_X Q_+^{(\tau)})}{2-\mu_X} + \mu_X Q_+^{(\tau)}\, log\, 2Q_+^{(\tau)}\right] \\ &+ \frac{1}{2}\left[(1-\mu_X Q_-^{(\tau)})\, log\, \frac{2(1-\mu_X Q_-^{(\tau)})}{2-\mu_X} + \mu_X Q_-^{(\tau)}\, log\, 2Q_-^{(\tau)}\right] \\ =& \frac{1}{2}\left[(1-\mu_X(1-\mu_Z))\, log\, \frac{2(1-\mu_X(1-\mu_Z))}{2-\mu_X} + \mu_X(1-\mu_Z)\, log\, 2(1-\mu_Z)\right] \\ &+ \frac{1}{2}\left[(1-\mu_X\mu_Z)\, log\, \frac{2(1-\mu_X\mu_Z)}{2-\mu_X} + \mu_X\mu_Z\, log\, 2\mu_Z\right]. \end{aligned}$$

When $\nu = |\tau - t_Z| = 2$, for example $\tau = 3, 7$ in Figure (7.6), we have that

$$Q_+^{(\tau)} = (1-\mu_Z)^2 + \mu_Z^2 \text{ and } Q_-^{(\tau)} = 2\mu_Z(1-\mu_Z).$$

And if we put these values in equation (7.10), the equation also becomes dependent on μ_Z. In the same way, when $\nu = 3$, for example $\tau = 2, 8$ in Figure (7.6), then

$$Q_+^{(\tau)} = (1-\mu_Z)^3 + 3\mu_Z^2(1-\mu_Z) \text{ and } Q_-^{(\tau)} = \mu_Z^3 + 3\mu_Z(1-\mu_Z)^2.$$

For $n_s = 2$, we can generalize this for any ν, where

$$Q_+^{(\tau)} = \sum_{k=0, k \text{ even}}^{\nu} \binom{\nu}{k} (1-\mu_Z)^{\nu-k} \mu_Z^k \quad \text{and}$$

$$Q_-^{(\tau)} = \sum_{k=1, k \text{ odd}}^{\nu} \binom{\nu}{k} (1-\mu_Z)^{\nu-k} \mu_Z^k,$$

so that $Q_+^{(\tau)} + Q_-^{(\tau)} = \sum_{k=0}^{\nu} \binom{\nu}{k} (1-\mu_Z)^{\nu-k} \mu_Z^k = (1 - \mu_Z + \mu_Z)^\nu = 1$ using the binomial theorem. Basically for $n_s = 2$ (thus also for simple model), $Q_+^{(\tau)}$ are the even terms and

$Q_{-}^{(\tau)}$ are the odd terms of the binomial theorem which depends on $\nu = |\tau - t_Z|$. This explains the varying values of $T_{GA}^{(\tau)}$ on the amended Ising model that is illustrated in Figures (6.16), (6.17) and (6.15).

Therefore as soon as $\tau \neq t_Z$, we have that $T_{ZX}^{(\tau)}$ becomes dependent on μ_Z. This is true for any $n_s \geq 2$ when $\mu_Z \neq \frac{n_s-1}{n_s}$. The value of $T_{ZX}^{(\tau)}$ does indeed increases as μ_X increases in the general model. The dependency on μ_Z values, comes into the Transfer Entropy values through different values of Q. Generally for $\mu_Z \neq \frac{n_s-1}{n_s}$, we get that $T_{ZX}^{(\tau)} \to T_{ZX}^{(t_Z)}$ as $\tau \to t_Z$ similar to the situation in Figure (7.6).

7.3 Cases of the general model

We have seen that the static probabilities stays uniform if we initialize with uniform and independent probabilities. If we want the transition probability of Z to stay uniform we need $\mu_Z = \frac{n_s-1}{n_s}$ and if we want the transition probability of X and Y to be uniform (with the exception of Ω influence) then let $\mu_X = \mu_Y = \frac{n_s-1}{n_s}$. However for the rest of this chapter

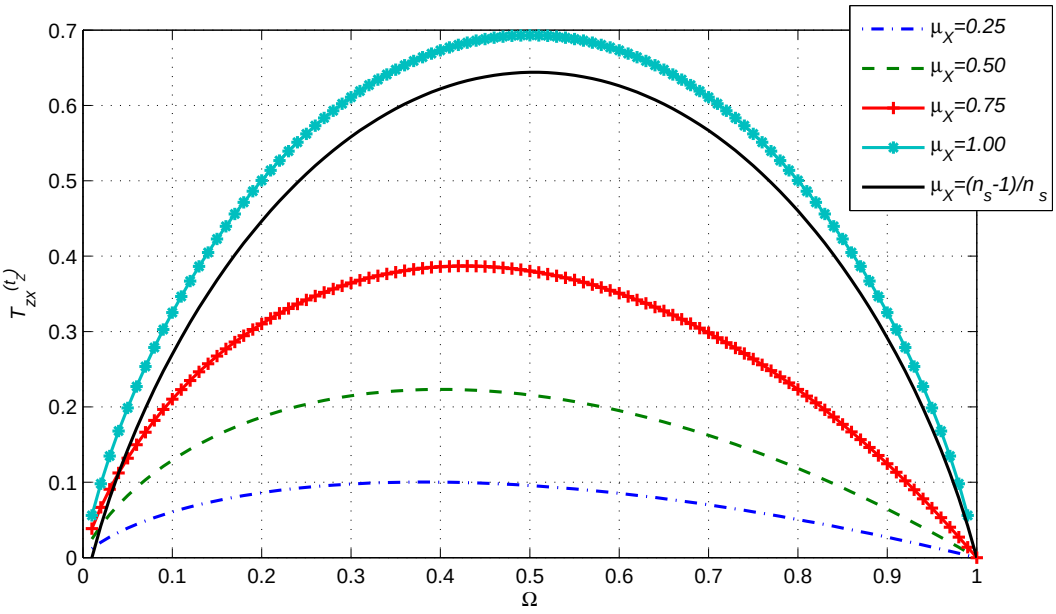

Figure 7.7: Analytical $T_{ZX}^{(t_Z)}$ versus Ω of equation (7.9) for different μ_X values. The maximum value $T_{ZX}^{(t_Z)} = log(2) \approx 0.6931$ is obtained when $\mu_X = 1$ and $\Omega = \frac{1}{2}$.

we will focus on Transfer Entropy values at t_Z, $T_{ZX}^{(t_Z)}$, which is independent of μ_Z. Figure (7.7) is equation (7.9) plotted over various Ω values. In Figure (7.7) when $\mu_X = 1$, $T_{ZX}^{(t_Z)}$ values are symmetrical over Ω as opposed to the other displayed values of μ_X where it is slightly skewed. When $\mu_X = 1$, the first term of equation (7.9) becomes 0, leaving us with a function that is symmetric over Ω. One can see that Transfer Entropy value is highest when $\Omega = 0.5$, the maximum value being $log(2) \approx 0.6931$. When $\mu_X = \frac{n_s-1}{n_s}$ (in this case n_s values ranging from 2 to 100 is plotted) the Transfer Entropy values approaches that of $\mu_X = 1$ as more and more values of different n_s is plotted. This is because $\mu_x = \frac{n_s-1}{n_s} \to 1$ as $n_s \to \infty$. It is worth pointing out that at $\Omega = 0$ and $\Omega = 1$, $T_{ZX}^{(t_Z)} = 0$. The former is due to the fact that X is not allowed to change at all hence becoming a constant. The latter is because when $\Omega = 1$, the transition probabilities of X becomes dependent on μ_X only with no extra restriction from Z and X becomes completely independent of Z. Thus in both cases Transfer Entropy correctly indicates independence between X and Z. We will further investigate the general model in terms of varying the value of Ω

Substituting $\mu_X = \frac{n_s-1}{n_s}$ into equation (7.9) makes $T_{ZX}^{(t_Z)}$ dependent on n_s and Ω such that

$$\begin{aligned} T_{ZX}^{(t_Z)} =& (1-\mu_X)\Omega\, log\, \frac{1-\mu_X}{1-\mu_X\Omega} + \mu_X\Omega\, log\, \frac{1}{\Omega} + (1-\Omega)\, log\, \frac{1}{1-\mu_X\Omega} \\ =& \frac{\Omega}{n_s}\, log\, \frac{\frac{1}{n_s}}{1-(\frac{n_s-1}{n_s})\Omega} + \frac{n_s-1}{n_s}\Omega\, log\, \frac{1}{\Omega} + (1-\Omega)\, log\, \frac{1}{1-(\frac{n_s-1}{n_s})\Omega} \\ =& \frac{\Omega}{n_s}\, log\, \frac{1}{n_s-(n_s-1)\Omega} + \frac{n_s-1}{n_s}\Omega\, log\, \frac{1}{\Omega} + (1-\Omega)\, log\, \frac{n_s}{n_s-(n_s-1)\Omega}. \end{aligned} \tag{7.11}$$

We shall see that Ω can be a function of n_s thus $T_{ZX}^{(t_Z)}$ can be made completely dependent on the number of states n_s.

7.3.1 Case 1: $\Omega = P(Z_{n-t_Z} = 1)$

Setting $\Omega = P(Z_{n-t_Z} = 1) = \frac{1}{n_s}$ means imposing the same condition as previously used, namely that X_n and Y_n can only change if $Z_n = 1$. However, as n_s increases, $Z_{n-t_Z} = 1$ becomes more and more restrictive (the states of X and Y becomes less and less able to change) since $\Omega = P(Z_{n-t_Z} = 1) = \frac{1}{n_s}$ gets smaller and smaller. Substituting $\Omega = \frac{1}{n_s}$ in

equation (7.9) renders

$$\begin{aligned} T_{ZX}^{(t_Z)} =&(1-\mu_X)\Omega\, log\, \frac{1-\mu_X}{1-\mu_X\Omega} + \mu_X\Omega\, log\, \frac{1}{\Omega} + (1-\Omega)\, log\, \frac{1}{1-\mu_X\Omega} \\ =&\frac{1-\mu_X}{n_s}\, log\, \frac{n_s(1-\mu_X)}{n_s-\mu_X} + \frac{\mu_X}{n_s}\, log\, n_s + \frac{n_s-1}{n_s}\, log\, \frac{n_s}{n_s-\mu_X} \\ =&log\, n_s + \frac{1}{n_s}\left[(1-\mu_X)\, log\,(1-\mu_X) - (n_s-\mu_X)\, log\,(n_s-\mu_X)\right]. \end{aligned} \tag{7.12}$$

This equation is illustrated in Figure (7.8) for various values of μ_X. Furthermore if we

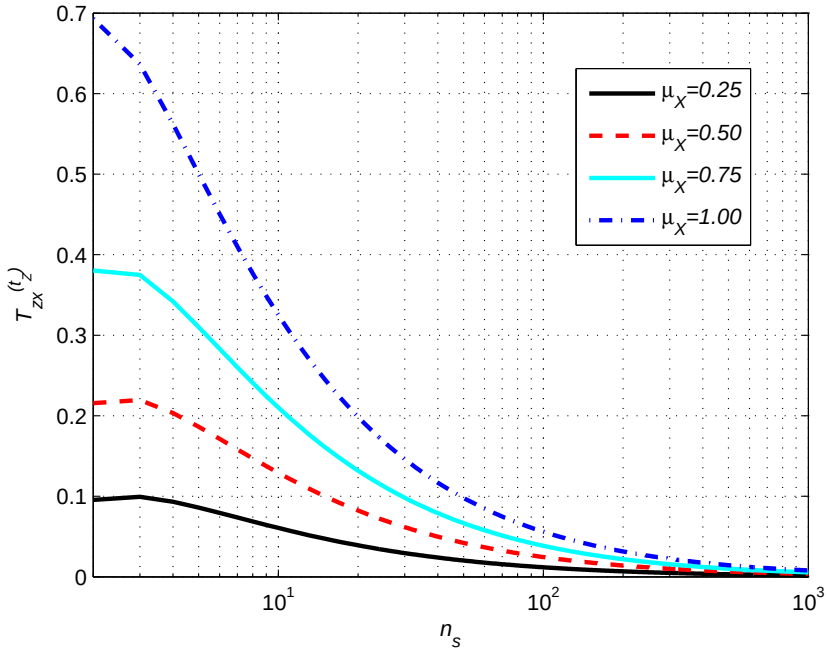

Figure 7.8: Analytical $T_{ZX}^{(t_Z)}$ versus n_s in equation (7.12) for Case 1.

substitute $\mu_X = \frac{n_s-1}{n_s}$ in equation (7.12) then we get

$$T_{ZX}^{(t_Z)} = log\, n_s + \frac{1}{n_s}\left[\frac{1}{n_s}\, log\, \frac{1}{n_s} - \frac{n_s(n_s-1)+1}{n_s}\, log\, \frac{n_s(n_s-1)+1}{n_s}\right] \tag{7.13}$$

in line with equation (7.11) and the equation is completely dependent on n_s. Figure (7.8) shows that $T_{ZX}^{(t_Z)} \to 0$ as $n_s \to \infty$ since $\Omega = \frac{1}{n_s} \to 0$ as $n_s \to \infty$. When this happens the condition on Z becomes so strict that X and Y can barely change thus practically becoming constants.

7.3.2 Case 2: $\Omega = P(Z_{n-t_Z} \neq 1)$

This case is the opposite of Case 1, since $\Omega = P(Z_{n-t_Z} \neq 1)$ and consequently X_n and Y_n can freely change as long as $Z_n \neq 1$ i.e the change is only restricted if $Z_n = 1$. This condition gets less strict (the processes becomes less and less dependent on Z) as n_s get bigger since $\Omega = \frac{n_s-1}{n_s}$ gets closer and closer to 1. The difference lay in the transition

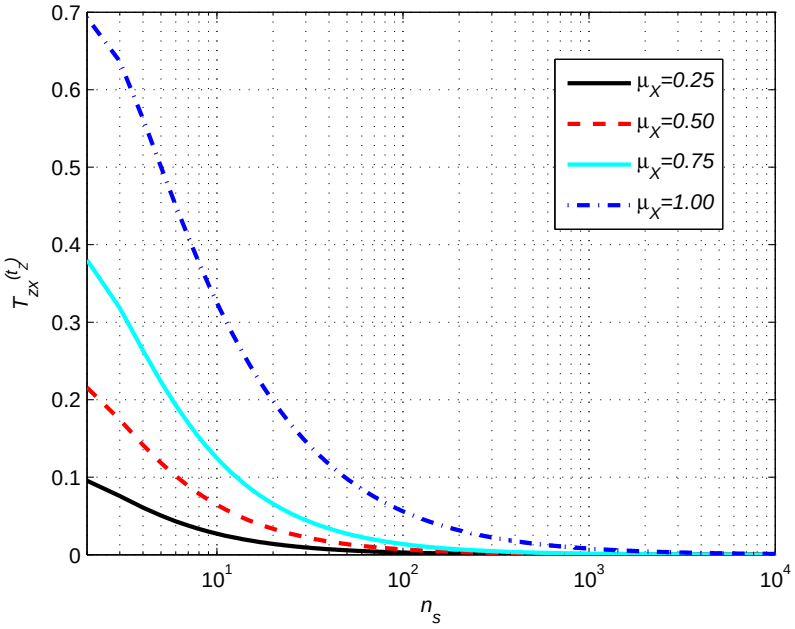

Figure 7.9: Analytical $T_{ZX}^{(t_Z)}$ versus n_s in equation (7.14) for Case 2.

probabilities of X and Y, since $\Omega = P(Z_{n-t_Z} \neq 1) = \frac{n_s-1}{n_s}$ and not $\frac{1}{n_s}$ (as in Case 1), so that equation (7.9) is

$$\begin{aligned} T_{ZX}^{(t_Z)} =& (1-\mu_X)\Omega \, log \, \frac{1-\mu_X}{1-\mu_X\Omega} + \mu_X\Omega \, log \, \frac{1}{\Omega} + (1-\Omega) \, log \, \frac{1}{1-\mu_X\Omega} \\ =& (1-\mu_X)\frac{n_s-1}{n_s} \, log \, \frac{n_s(1-\mu_X)}{n_s-\mu_X(n_s-1)} + \frac{\mu_X(n_s-1)}{n_s} \, log \, \frac{n_s}{n_s-1} \\ &+ \frac{1}{n_s} \, log \, \frac{n_s}{n_s-\mu_X(n_s-1)} \\ =& log \, n_s + \frac{(n_s-1)}{n_s}(1-\mu_X) \, log \, (1-\mu_X) - \frac{\mu_X}{n_s}(n_s-1) \, log \, (n_s-1) \\ &- \frac{(n_s(1-\mu_X)+\mu_X)}{n_s} \, log \, (n_s(1-\mu_X)+\mu_X). \end{aligned} \tag{7.14}$$

Figure (7.9) where equation (7.14) is plotted for various μ_X shows that since $\Omega = \frac{n_s-1}{n_s}$ we have that $\Omega \to 1$ as $n_s \to \infty$. $\Omega = 1$ implies that the condition is fulfilled all the time therefore there is no dependence on Z anymore. Thus in this model X, Y and Z becomes more and independent as $n_s \to \infty$ and therefore $T_{ZX}^{(t_Z)} \to 0$ as $n_s \to \infty$ as in Case 1.

Substituting $\mu_X = \frac{n_s-1}{n_s}$ in equation (7.14) we get that

$$\begin{aligned} T_{ZX}^{(t_Z)} =& log\, n_s + \frac{(n_s-1)}{n_s}\frac{1}{n_s}\, log\, \frac{1}{n_s} - \frac{(n_s-1)}{n_s^2}(n_s-1)\, log\, (n_s-1) \\ & - \frac{(n_s\frac{1}{n_s} + \frac{(n_s-1)}{n_s})}{n_s}\, log\, \left(n_s\frac{1}{n_s} + \frac{(n_s-1)}{n_s}\right) \\ =& log\, n_s + \frac{(n_s-1)}{n_s^2}\, log\, \frac{1}{n_s} - \frac{(n_s-1)^2}{n_s^2}\, log\, (n_s-1) - \frac{2n_s-1}{n_s^2}\, log\, \left(\frac{2n_s-1}{n_s^2}\right). \end{aligned} \tag{7.15}$$

7.3.3 Case 3: $\Omega = \frac{1}{2}$

Figure (7.10) is equation (7.8) plotted for $\Omega = \frac{1}{2}$ over different n_s for various μ_X values. This can be achieved in simulations by setting the condition so that it is fulfilled by half of the possible state space all the time so that $\Omega = \frac{\frac{n_s}{2}}{n_s} = \frac{1}{2}$. One can clearly see in Figure (7.10) that $T_{ZX}^{(t_Z)}$ becomes n_s independent and only depends on μ_X. This can be understood by substituting $\Omega = \frac{1}{2}$ in equation (7.9) so that

$$\begin{aligned} T_{ZX}^{(t_Z)} =& (1-\mu_X)\Omega\, log\, \frac{1-\mu_X}{1-\mu_X\Omega} + \mu_X\Omega\, log\, \frac{1}{\Omega} + (1-\Omega)\, log\, \frac{1}{1-\mu_X\Omega} \\ =& \frac{1-\mu_X}{2}\, log\, \frac{2(1-\mu_X)}{2-\mu_X} + \frac{\mu_X}{2}\, log\, 2 + \frac{1}{2}\, log\, \frac{2}{2-\mu_X} \\ =& log\, 2 + \frac{1}{2}\left[(1-\mu_X)\, log\, (1-\mu_X) - (2-\mu_X)\, log\, (2-\mu_X)\right]. \end{aligned} \tag{7.16}$$

As we have seen before, when we let $\Omega = \frac{1}{2}$ we get a the simple model equation (7.1).

However, we do get a dependence on n_s (in line with equation (7.11)) if $\mu_X = \frac{n_s-1}{n_s}$ is substituted in equation (7.16) so that

$$\begin{aligned} T_{ZX}^{(t_Z)} &= log\, 2 + \frac{1}{2}\left[\frac{1}{n_s}\, log\, \frac{1}{n_s} - \frac{n_s+1}{n_s}\, log\, \frac{n_s+1}{n_s}\right] \\ &= log\, 2 + \frac{1}{2}\, log\, n_s - \frac{n_s+1}{2n_s}\, log\, (n_s+1). \end{aligned} \tag{7.17}$$

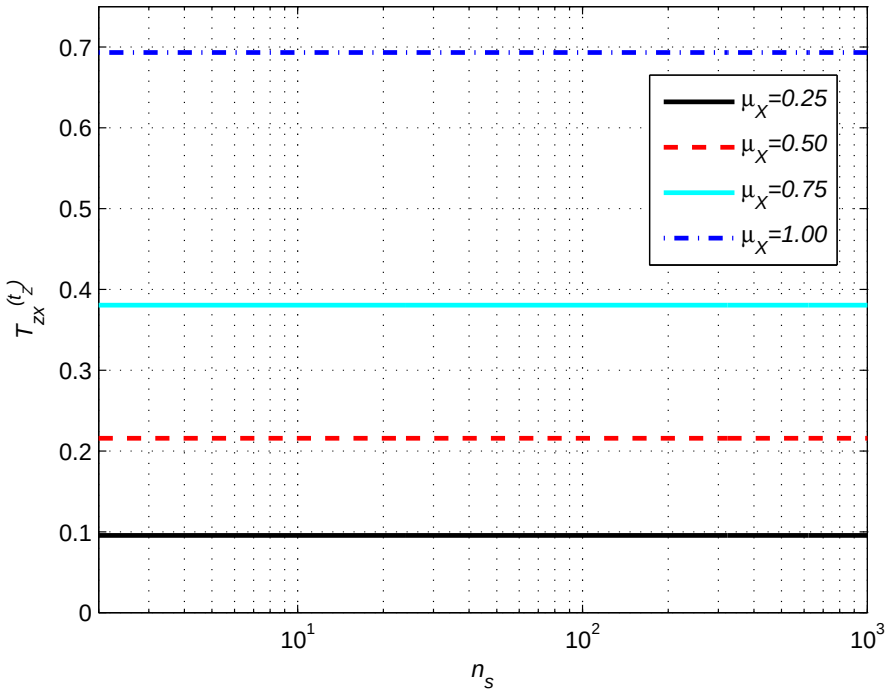

Figure 7.10: Analytical $T_{ZX}^{(t_Z)}$ versus n_s in equation (7.8) for Case 3. Transfer Entropy values are independent of n_s and completely dependent on μ_X akin to the simple model.

In equation (7.17) one can see that even if initially there is some independence on n_s, the $T_{ZX}^{(t_Z)}$ value in converges quite rapidly to $log\,(2)$ as $n_s \to \infty$, which is the value of $T_{ZX}^{(t_Z)}$ with $\mu_X = 1$ of equation (7.8) as seen in Figure (7.10).

7.3.4 Discussion

For any restriction that we place on Z_{n-t_Z}, it is $\Omega = P(\text{ condition fulfilled })$, the probability of the fulfilling the condition that matters. The different cases highlight the fact that there can be different types of restrictions (conditions) that will affect the values of Transfer Entropy in different ways. In Case 1 where $\Omega = P(Z_{n-1} = 1)$ and Case 2 where $\Omega = P(Z_{n-1} \neq 1)$, for $n_s = 2$ they both become the simple model and thus a type of Case 3. This is due to the fact that $n_s = 2$ leads to $\Omega = \frac{1}{2}$ since $P(Z_{n-1} = 1) = P(Z_{n-1} \neq 1) = \frac{1}{2}$. Basically for $n_s = 2$, all the cases are indistinguishable. Figure (7.11) plots

equations (7.13), (7.15) and (7.17) where $\mu_X = \frac{n_s-1}{n_s}$ is substituted into $T_{ZX}^{(t_Z)}$ making the transition probability of X uniform save for the influence of Ω for Case 1, Case 2 and Case 3 respectively. The figure showcases the asymptotical behaviour of each case.

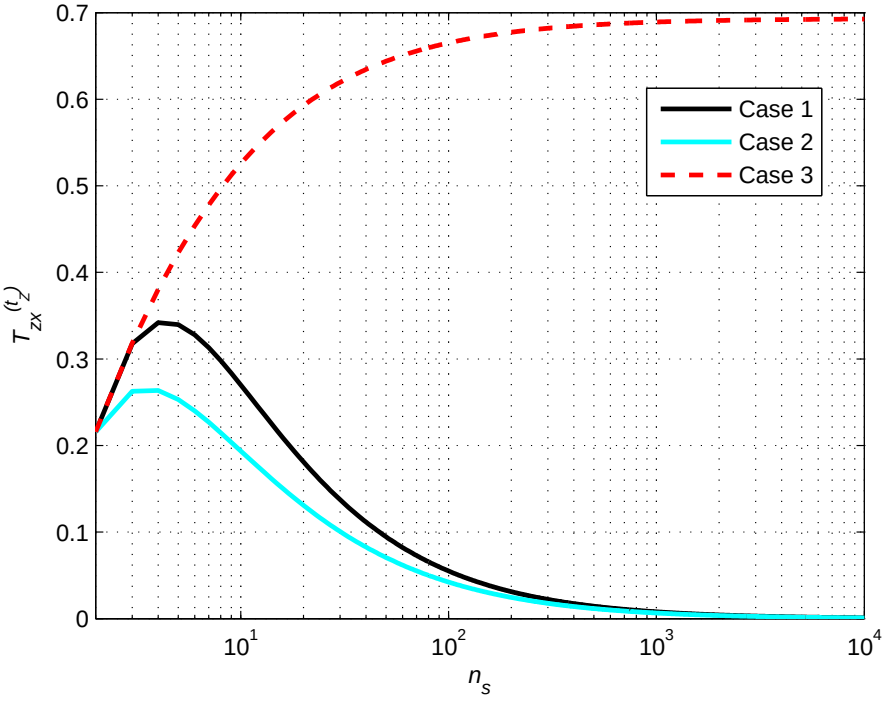

Figure 7.11: Analytical $T_{ZX}^{(t_Z)}$ versus n_s with $\mu_X = \frac{n_s-1}{n_s}$ in equations (7.13), (7.15) and (7.17) for Cases $1-3$ respectively. Case 1 and 2 approaches 0. Case 3 approaches $log(2)$.

What we have illustrated are two extreme cases in the form of Case 1 and Case 2 and a middle ground in the form of Case 3 . For Case 1 , $\Omega = \frac{1}{n_s}$ and thus $\Omega \to 0$ as $n_s \to \infty$. For Case 2 , $\Omega = \frac{n_s-1}{n_s}$ and thus $\Omega \to 1$ as $n_s \to \infty$. This makes sense because $\Omega = 0$ simply implies that X and Y are not allowed to change thus becoming constants and $\Omega = 1$ implies that the condition is always fulfilled making X and Y just like Z, randomly assigned values depending only on μ_X and μ_Y respectively with no restrictions or driving factor. For Case 3 however, $T_{ZX}^{(t_Z)}$ stabilizes to constant $log\,(2) \approx 0.6931$ as $n_s \to \infty$. This asymptotic behaviour can be attributed to $\mu_X = \frac{n_s-1}{n_s} \to 1$ as $n_s \to \infty$. Therefore, theoretically the Transfer Entropy does indeed captures the relationship between these stochastic processes.

Chapter Summary

From the analytical values of this model we see that Transfer Entropy successfully detects the 'dependency' of different restrictions that was implanted in the model. While covariance, Mutual Information and conditional Mutual Information fails to detect anything and gives 0, the Transfer Entropy clearly gives us nonzero values that can be taken as an indication of 'causality'. Ω stands for the percentage of states of Z that allows changes in X and Y and serves as an indication of the level of restriction imposed by Z on X and Y. The experiments with different values of μ_X and Ω, highlights how the different magnitudes of $T_{ZX}^{(\tau)}$ reflects values of μ_X, the intrinsic probability that X will change regardless of outsides influence, as well as the values of Ω, representing the outside influence on X.

Furthermore the variable $Q_{sgn(\gamma)}^{(\tau)}$ which represents the probability of the condition being fulfilled given the current information available time τ, enables us to understand how μ_Z influences $T_{ZX}^{(\tau)}$ so that $T_{ZX}^{(\tau)} \neq 0$ even when $\tau \neq t_Z$. This shows the importance of testing for different causal lags, as only the largest value $T_{ZX}^{(\tau)}$ is the real lag. Therefore, the intrinsic probabilities of both causal and effected processes are also very important in determining the values of Transfer Entropy between them and should be taken into account whenever one is trying to make sense of the different magnitudes of Transfer Entropy. More importantly, we have proved that using the theoretical value of Transfer Entropy it is possible to pinpoint the exact time lag in which this 'causal' connection occurs in this analytically solvable model. We now proceed to simulations.

Chapter 8

Finite sampling effects and estimations

We have shown that theoretically, Transfer Entropy has the potential to detect the causal lag involved in the imposed causal relationship on the model. The next step is to simulate the model to develop some experience for how sample sizes (i.e. limited data sets) will influence the behaviour of the Transfer Entropy as we increase the number of states n_s. We simulate the model in MATLAB by generating stochastic processes X, Y and Z with sample size S. We also simulated a null model to further illustrate these finite sampling effects as well as some proposed corrections along the lines of significant testing. Furthermore we discuss some popular methods of entropy estimation in relation to applying Transfer Entropy on real data sets and how we decided to use the most common classical histogram method.

8.1 Simulation of the toy model

As we increase the number of states n_s, we will need to increase the simulated data required to get accurate probabilities and the effect will be evident in the simulation. We will see the effects of different sample sizes (length of each stochastic process) in probability estimation. Recall that the Transfer Entropy definition in equation (4.5) is $T_{YX}^{(\tau)} = E\left[log\,\frac{P(X|X^{-1},Y^{-\tau})}{P(X|X^{-1})}\right]$. The Transfer Entropy values displayed in this section are the product of applying equation (4.5) on the simulated data of the toy model.

8.1.1 Simulation of $n_s = 2$

We simulate $n_s = 2$ with $t_Z = 5$. The usage of sample size $S = 10000$ for $n_s = 2$ appears to give sufficient statistics since Figures (8.1) and (8.2) look identical to their analytical counterparts Figures (7.5) and (7.6).

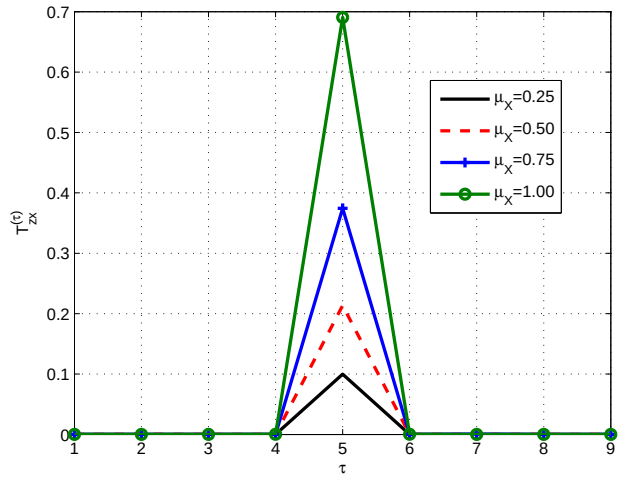

Figure 8.1: $T^{(\tau)}_{ZX}$ values obtained using equation (4.5) on simulated toy model with fixed values of $\mu_Z = \frac{1}{2}$, $t_Z = 5$ and $n_s = 2$. Simulated version of Figure (7.5).

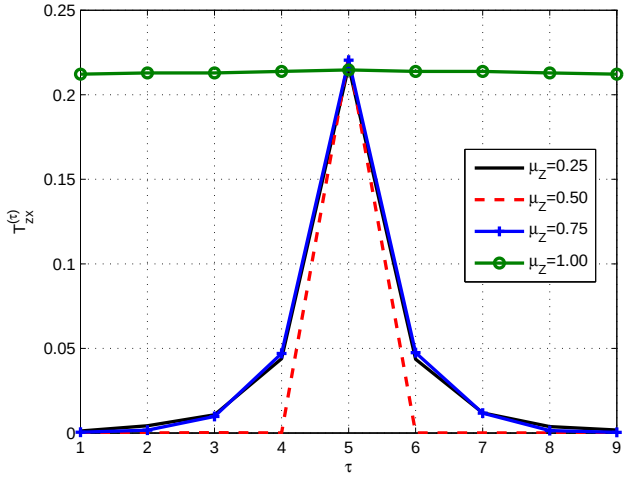

Figure 8.2: $T^{(\tau)}_{ZX}$ values obtained using equation (4.5) on simulated toy model with fixed values of $\mu_X = \frac{1}{2}$, $t_Z = 5$ and $n_s = 2$. Simulated version of Figure (7.6).

To illustrate the effect of sample size, Figures (8.3) and (8.4) depicts Transfer Entropy between X and Z in both directions when $\mu_X = \mu_Z = \frac{1}{2}$ for sample sizes $S = 10000$ and $S = 100$ respectively. We know from equation (7.10), that $T_{ZX}^{(5)} \approx \frac{3}{2} log\, 2 - \frac{3}{4} log\, 3 \approx 0.2158$ and by definition all the other Transfer Entropy values in both direction are supposed to be 0.

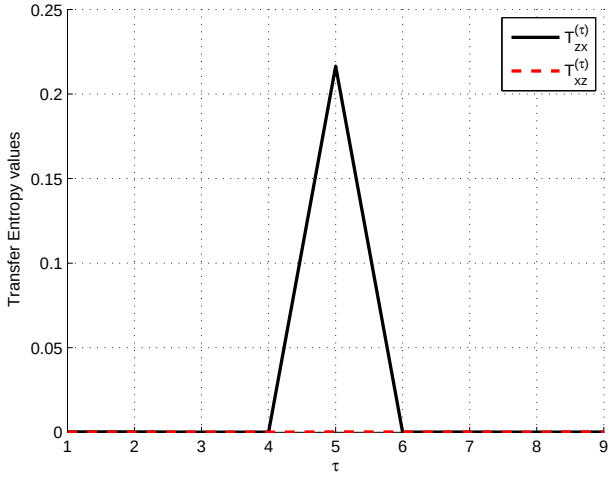

Figure 8.3: $T_{ZX}^{(\tau)}$ and $T_{XZ}^{(\tau)}$ values obtained using equation (4.5) on simulated toy model with $n_s = 2$, $t_Z = 5$, $\mu_X = \mu_Z = \frac{1}{2}$ and sample size $S = 10000$.

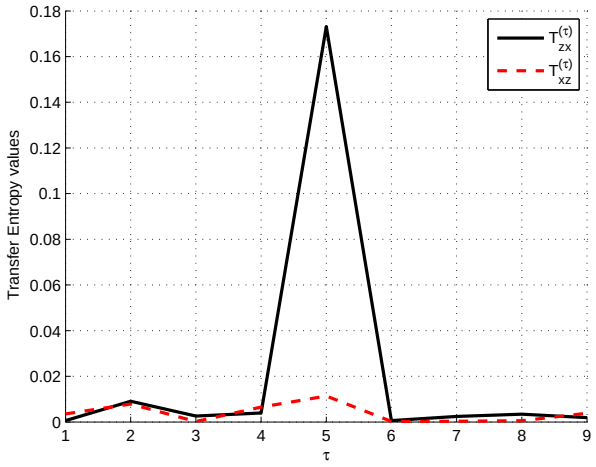

Figure 8.4: $T_{ZX}^{(\tau)}$ and $T_{XZ}^{(\tau)}$ values obtained using equation (4.5) on simulated toy model with $n_s = 2$, $t_Z = 5$, $\mu_X = \mu_Z = \frac{1}{2}$ and sample size $S = 100$.

The values of $T_{ZX}^{(5)}$ stands out in both Figures (8.3) and (8.4), thus correctly indicating the implanted causal lag at $t_Z = 5$. However in Figure (8.4), one can see that the value $T_{ZX}^{(5)}$ is now much further from the theoretical value of 0.2158 and $T_{ZX}^{(\tau)}$ for $\tau \neq 5$ values are not as close to 0 as Figure (8.4). We attribute this to the lack of statistics or in other words insufficient data points to get the actual probabilities.

8.1.2 Simulation of Case 3

In this subsection we measure the Transfer Entropy values of simulated Case 3 and compare them to the theoretical values in subsection (7.3.3). In simulation of Case 3, the value of $\Omega = \frac{1}{2}$ can be replicated by making $\Omega \approx \frac{\frac{n_s}{2}}{n_s}$ such that the condition is fulfilled by approximately half of the states all the time. We set the sample size, $S = 10000$ and $t_z = 5$ for illustration purposes. In Figures (8.5) and (8.6), Transfer Entropy applied on simulated

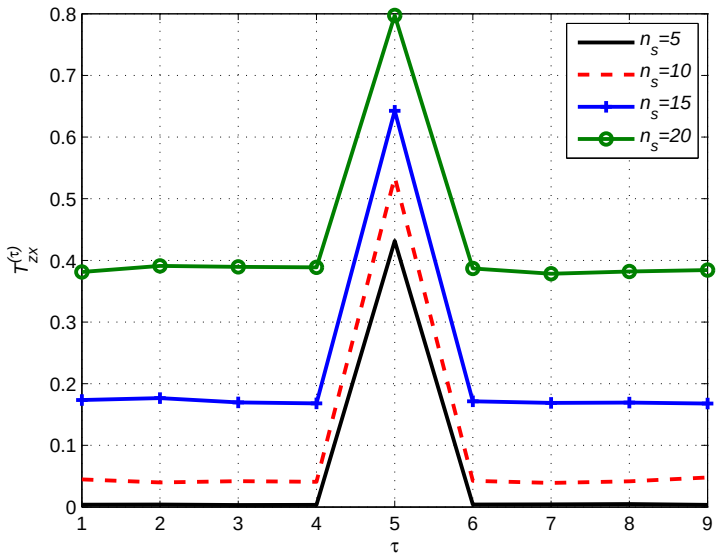

Figure 8.5: $T_{ZX}^{(\tau)}$ obtained using equation (4.5) on simulated toy model of Case 3 with $\mu_X = \frac{n_s-1}{n_s}$ and $\mu_Z = \frac{n_s-1}{n_s}$ for $n_s = 5, 10, 15, 20$. The only nonzero analytical values (correct the 4 decimal places) of peaks at $T_{ZX}^{(5)}$ given by equation (7.17) are 0.4228, 0.5256, 0.5685 and 0.5926 respectively.

data of Case 3 is plotted. The theoretical values of $T_{ZX}^{(5)}$ in Figure (8.5) can be obtained by substituting the appropriate n_s value into equation (7.17). One can clearly see that the

larger n_s gets the more inaccurate it becomes. Some of the peak values are clearly different from its theoretical value since they are larger than $log(2) \approx 0.6931$ whereas the values of equation (7.17) would be approaching $log(2)$ from below for $n_s \to \infty$ as seen in Figure (7.11). When $\tau \neq t_Z$, the values of $T_{ZX}^{(\tau)}$ for $\mu_Z = \frac{n_s-1}{n_s}$ are theoretically 0.

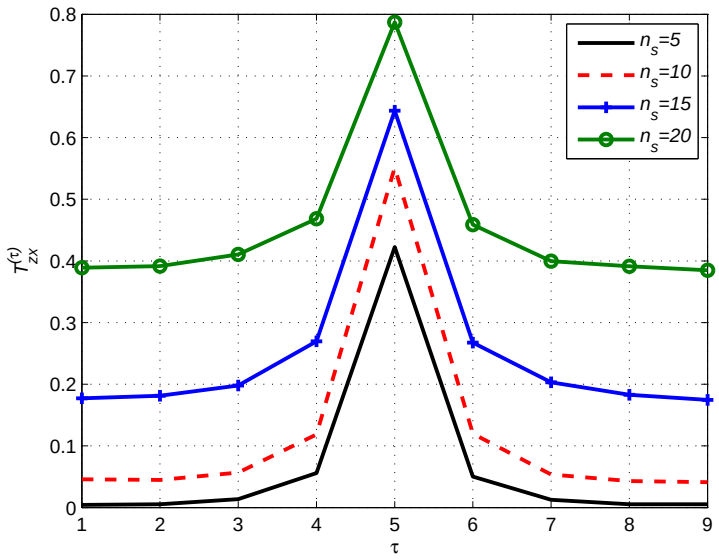

Figure 8.6: $T_{ZX}^{(\tau)}$ obtained using equation (4.5) on simulated toy model of Case 3 with $\mu_X = \frac{n_s-1}{n_s}$ and $\mu_Z = \frac{1}{2}$ for $n_s = 5, 10, 15, 20$. The analytical values (correct the 4 decimal places) of peaks at $T_{ZX}^{(5)}$ given by equation (7.16) are also 0.4228, 0.5256, 0.5685 and 0.5926 respectively.

In Figure (8.6), we have $\mu_Z = \frac{1}{2}$ instead of $\mu_Z = \frac{n_s-1}{n_s}$ in Figure (8.5). This is the only difference between the two figures. Recall from discussions in subsection (7.2.3) that the values of $T_{ZX}^{(t_Z)}$ are not influenced by μ_Z, therefore at $t_Z = 5$ the theoretical values of $T_{ZX}^{(t_Z)}$ of Figures (8.6) and (8.5) are exactly the same. This can be verified for each n_s by substituting $\mu_X = \frac{n_s-1}{n_s}$ into equation (7.16). However, the figures differ for values of $T_{ZX}^{(\tau)}, \tau \neq t_Z$. Also discussed in subsection (7.2.3), is the fact that when $\mu_Z \neq \frac{n_s-1}{n_s}$ as in Figure (8.6), there exist nonzero values for $T_{ZX}^{(\tau)}, \tau \neq t_Z$ which is influenced by μ_Z through values of Q. The influence of Q can be understood by examining equation (7.8) that is the general equation for Transfer Entropy on the toy model in which both equations (7.17) and (7.16) are derived from.

8.1.3 Different cases of the general model

We proceed to investigate how much the sample size affects the different cases on the general model. In this subsection we plot values of Transfer Entropy $T_{ZX}^{(t_Z)}$ with $\mu_Z = \frac{n_s-1}{n_s}$. Without loss of generality we utilize $t_Z = 1$ for all the simulations. Obviously the larger S is, the closer the approximation from simulated data sets are to the analytical values. In Figure (8.7), we plot the analytical value of equation (7.13) for each n_s alongside the simulations with varying lengths (sample sizes) of simulated data sets for Case 1 where the condition gets stricter as n_s gets bigger since $\Omega = \frac{1}{n_s} \to 0$. One can see that the approximation does not stray too far from the analytical value as opposed to other two cases. Case 2 is the opposite of Case 1 where the condition gets less and less strict which

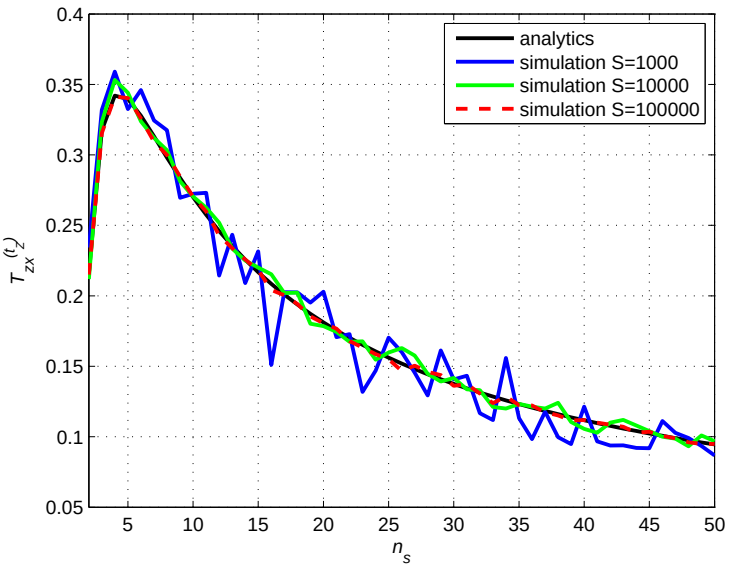

Figure 8.7: Transfer Entropy $T_{ZX}^{(t_Z)}$ versus number of state n_s for Case 1. Analytical values obtained from equation (7.13) and simulated values acquired using equation (4.5) on simulated data of varying sample size S.

is probably why some of the values seen in Figure (8.8) that are supposed to converge to 0 analytically, diverge instead. Figure (8.9) represents Case 3 where $\Omega = \frac{1}{2}$. One can see that some of the approximated values also diverge and does not converge to $log(2)$ as it is supposed to.

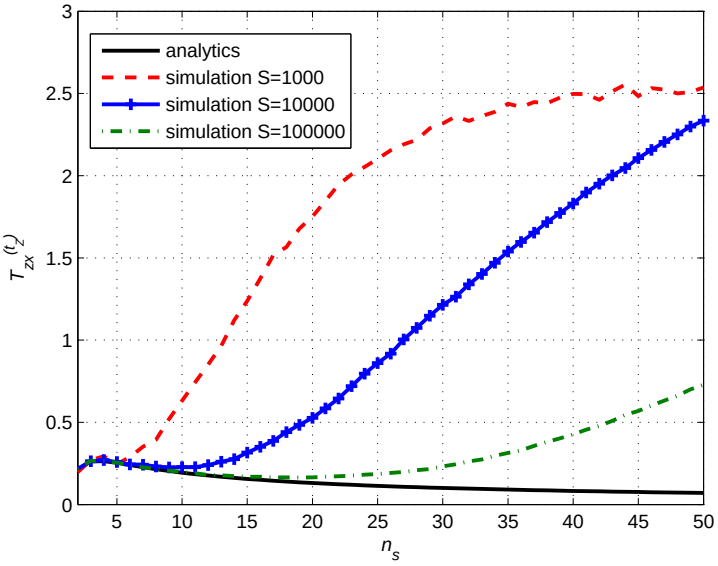

Figure 8.8: Transfer Entropy $T_{ZX}^{(t_Z)}$ versus number of state n_s for Case 2. Analytical values obtained from equation (7.15) and simulated values acquired using equation (4.5) on simulated data of varying sample size S.

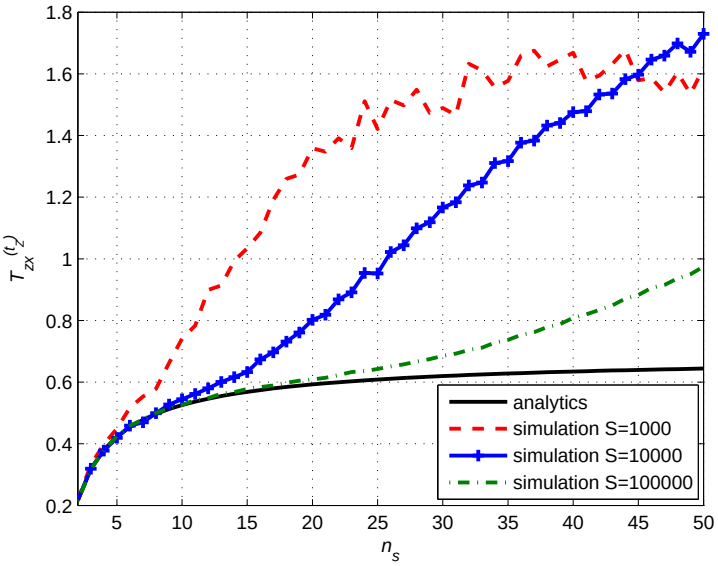

Figure 8.9: Transfer Entropy $T_{ZX}^{(t_Z)}$ versus number of state n_s for Case 3. Analytical values obtained from equation (7.17) and simulated values acquired using equation (4.5) on simulated data of varying sample size S.

8.2 The null model

To differentiate what is happening because of lack of statistics and what is really happening in the model, we look at a null model where no conditions are imposed. This null model can be described as the general model with $\Omega = 1$ where Z does not effect X or Y. Let the model be contained of stochastic processes X, Y and Z that can assume values in the set of states $A = \{1, \cdots, n_s\}$ at every time step $n = 1, ..., S$. Let μ_X, μ_Y and μ_Z be the internal (not influence by other processes) probabilities that the variables X, Y and Z changes at every time step respectively, so that the transition probabilities become

$$P(X_n = \alpha | X_{n-1} = \beta) = \begin{cases} 1 - \mu_X & \text{if } \alpha = \beta \\ \frac{1}{n_s - 1}\mu_X & \text{if } \alpha \neq \beta \end{cases}$$

$$P(Y_n = \alpha | Y_{n-1} = \beta) = \begin{cases} 1 - \mu_Y & \text{if } \alpha = \beta \\ \frac{1}{n_s - 1}\mu_Y & \text{if } \alpha \neq \beta. \end{cases}$$

and

$$P(Z_n = \alpha | Z_{n-1} = \beta) = \begin{cases} 1 - \mu_Z & \text{if } \alpha = \beta \\ \frac{1}{n_s - 1}\mu_Z & \text{if } \alpha \neq \beta. \end{cases}$$

8.2.1 Transfer Entropy on the null model

Due to independence, one would expect the transition probabilities between the processes to be independent of each other. For example, the transition probability between X and Z becomes

$$P(X_n = \alpha | X_{n-1} = \beta, Z_{n-\tau} = \gamma) = P(X_n = \alpha | X_{n-1} = \beta)$$

for any $\alpha, \beta, \gamma \in A$ and time step τ. Consequently, the ratio $\frac{P(X_n = \alpha | X_{n-1} = \beta, Z_{n-\tau} = \gamma)}{P(X_n = \alpha | X_{n-1} = \beta)}$ will always be 1 therefore the Transfer Entropy values should all be 0. Figures (8.10) and (8.11) display that the Transfer Entropy diverge from 0 for larger n_s especially for smaller sample sizes. The aim of doing this is to get some feel of what values of n_s are appropriate for the different sample sizes (data set lengths) that we have. We clearly see that approximations on insufficient sample size leads to spurious values.

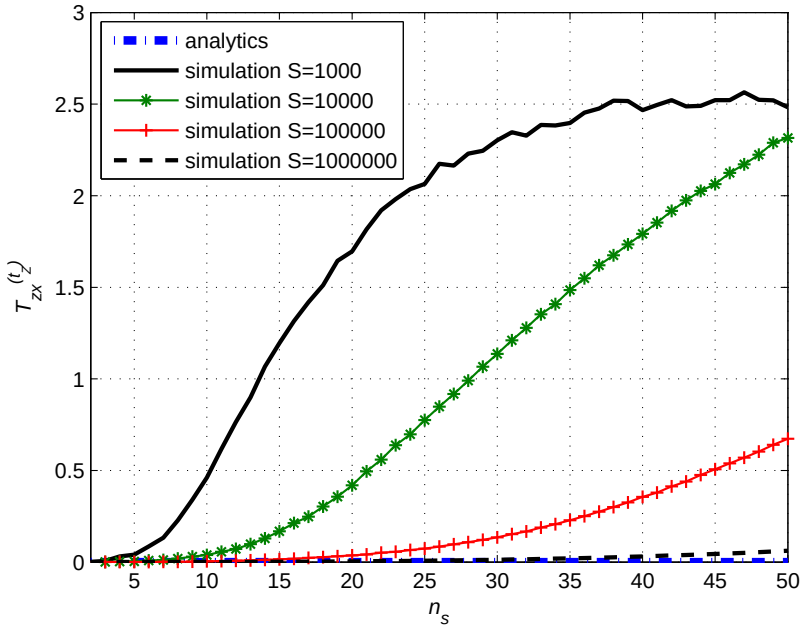

Figure 8.10: Transfer Entropy $T_{ZX}^{(t_Z)}$ versus number of state n_s for null model. Analytical values are 0 and simulated values acquired using equation (4.5) on simulated data of varying sample size S.

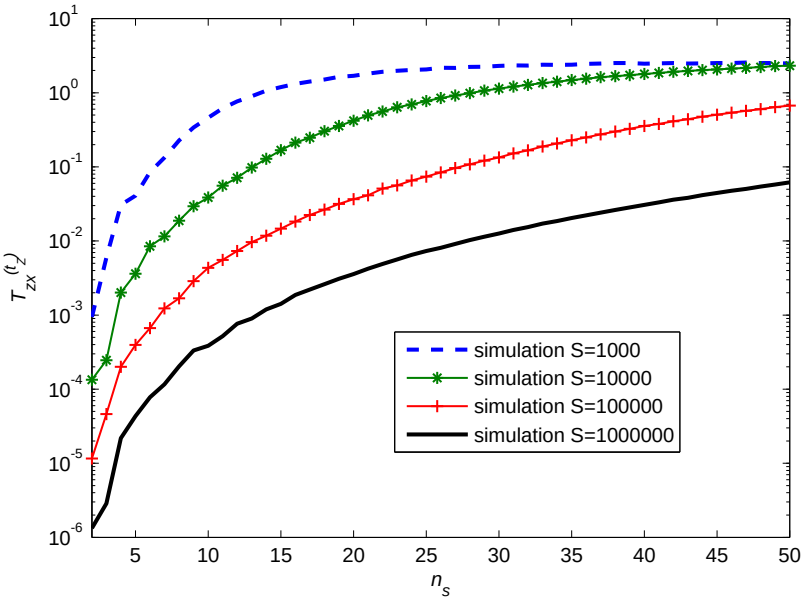

Figure 8.11: Figure (8.10) with log values on y axis

8.2.2 Mutual Information and covariance on the null model

We see that this problem is not only exclusive to Transfer Entropy. It applies to any case where probability needs to be estimated. We illustrate the situation for Mutual Information in Figures (8.12) and (8.13) . The covariance values are displayed in Figure (8.14).

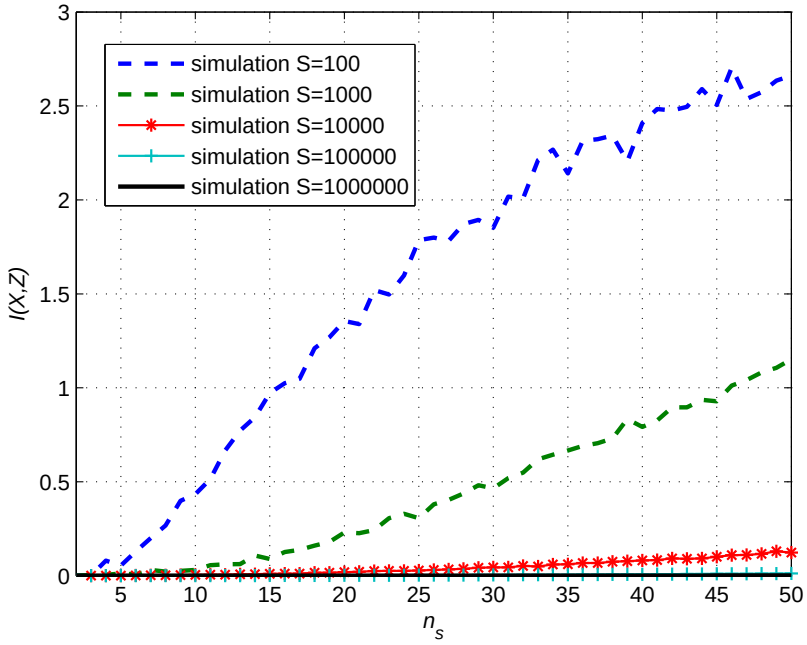

Figure 8.12: Mutual Information $I(X, Z)$ versus number of states n_s for null model. Analytical values are 0. Simulated values acquired using equation (2.12) on simulated data of varying sample size S.

8.3 Correcting for finite sampling effects

The observed existence of spurious detection or overestimation [74] is not uncommon and has been reported in relation to causality measures in [100, 51, 101, 79, 71]. These spurious values are caused by bias in relation to individual dynamics, state space reconstruction, coupling measure on so on so forth. The bias of an estimator is the difference between estimators expectation value and it's theoretical value. Bias in estimation causes non-zero spurious values when there is no causal effect and this problem is not only unique for Transfer Entropy [79]. This is a problem in which positive bias may be misinterpreted

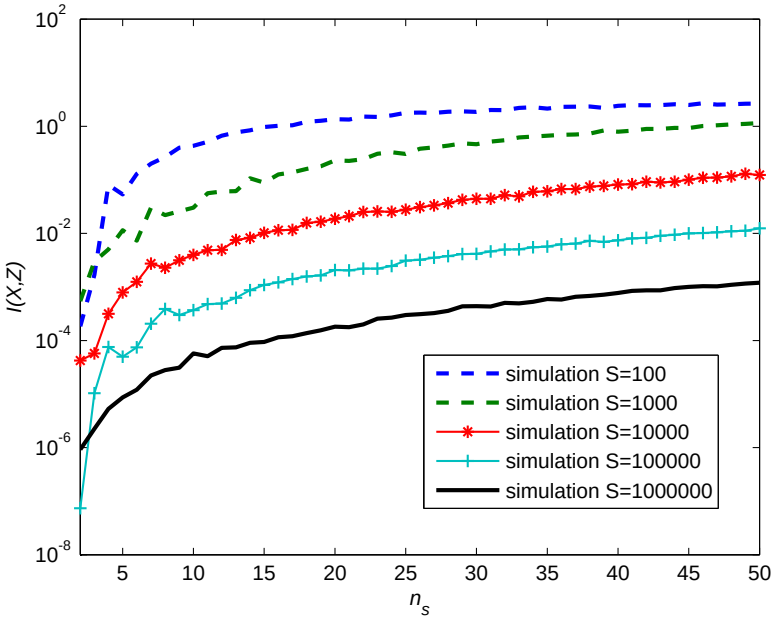

Figure 8.13: Figure (8.12) with log values on y axis

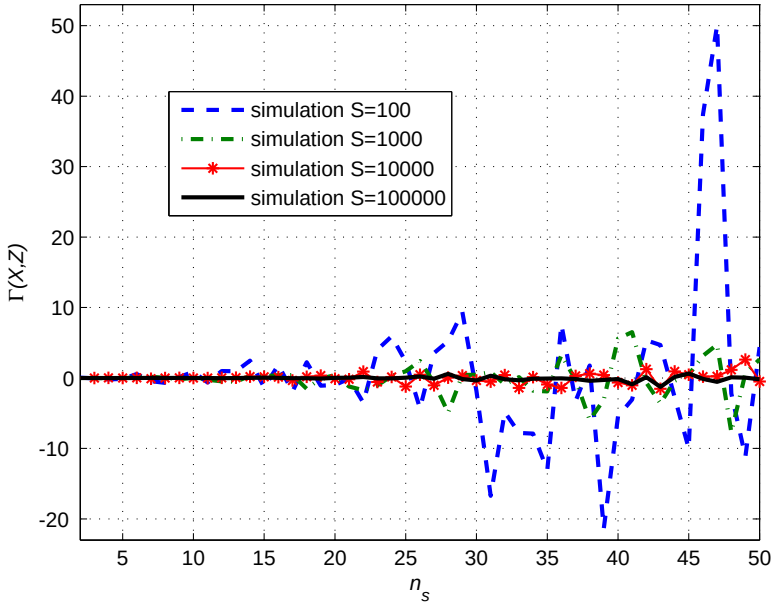

Figure 8.14: Covariance $\Gamma(X, Z)$ versus number of states n_s for null model. Analytical values are 0 and simulated values acquired using equation (1.2) on simulated data of varying sample size S.

as weak coupling when there is actually no causal effect. Therefore there needs to be way to indicate significance and reduce bias so that the causal measure gives zero values when there is no causal relationship. In [100, 51] correction terms to cancel out the bias related errors are suggested. Another alternative to cope with the finite sampling effects is significant testing. Surrogates have been suggested as a form of significant testing for Transfer Entropy [101, 102, 78, 75].

8.3.1 Surrogates for significant testing

When compared to G-causality, it is often pointed out that significant testing is not present for Transfer Entropy. Schreiber outlined that directionality can only be concluded if the value of Transfer Entropy is 0 in one direction and nonzero in another. However due to bias, the value 0 is not normally obtained for Transfer Entropy in real data sets. [79] points out the importance of having significant test for causality measures in terms avoiding false directionality conclusions.

[78] claims that the only practical significant testing for Transfer Entropy is probably in the form of surrogates. Surrogates data sets are synthetically generated data which ideally preserve all properties of the underlying system except the one being tested [101]. There are many different types of surrogates to serve different purposes. Fourier surrogates are used to randomize frequencies [90, 101]. Randomizing temporal values have been done using permutation surrogates [78], time shift test [102] and twin surrogates [101]. Surrogates have also been used in testing whether or not data sets are nonlinear [75]. Surrogates in the form of reshuffled time series are utilized in [72, 18]. The idea is to break the coupling (causal link) but maintain dynamics in hope that one can differentiate cause and effect from the any other dynamics.

Significant testing with surrogate is usually done as a standard one sided hypothesis test where the null hypothesis is that the two systems (time series) are independent. Attempts are made to reject the null hypothesis with a certain confidence level. A more inclusive test taking into account different directions and non-directionality is proposed in [79]. Rather than testing for surrogates separately it has also been suggested that significant testing can be done in a form of modified information theoretic functionals [85].

8.3.2 Effective and Corrected Transfer Entropy

From Figures (8.5) and (8.6), it seems like the values are simply shifted upwards and if one could simply subtract values related to the shift then perhaps the true values would be obtained. This is the idea behind the effective and corrected Transfer Entropy. Effective Transfer Entropy [71] between two time series is the modification of Transfer Entropy defined as the difference of Transfer Entropy computed on the original time series and Transfer Entropy computed between a surrogate time series where the driving process is randomly shuffled. Therefore in relation to our definition of Transfer Entropy in equation (4.5) the effective Transfer Entropy can be defined as

$$ET_{YX}^{(\tau)} = T_{YX}^{(\tau)} - T_{Y_S,X}^{(\tau)} \tag{8.1}$$

where Y_S is the randomly shuffled surrogate of time series Y. Figures (8.15) and (8.16) displays the values on effective Transfer Entropy on Case 3 of the general model and null model in direct contrast to Figures (8.9) and (8.10).

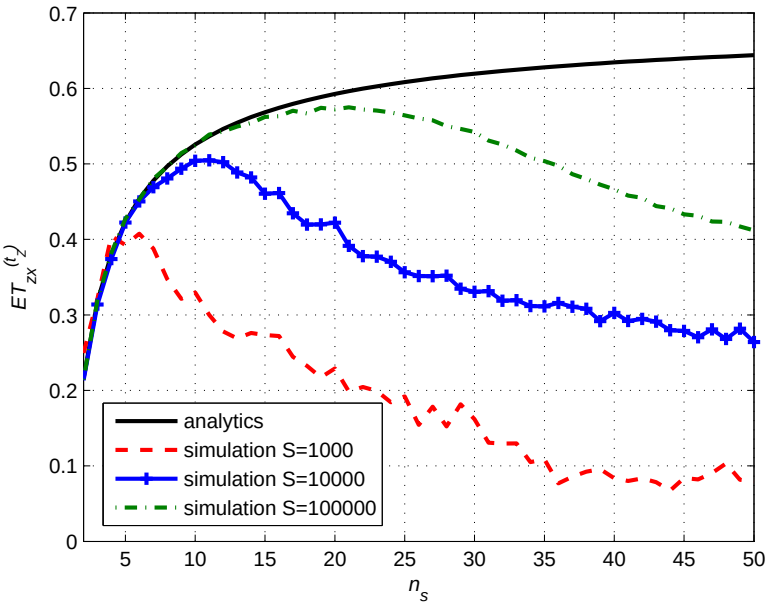

Figure 8.15: Effective Transfer Entropy $ET_{ZX}^{(t_Z)}$ versus number of states n_s for Case 3. Analytical values are obtained with equation (7.17) and simulated values acquired using equation (8.1) on simulated data of varying sample size S.

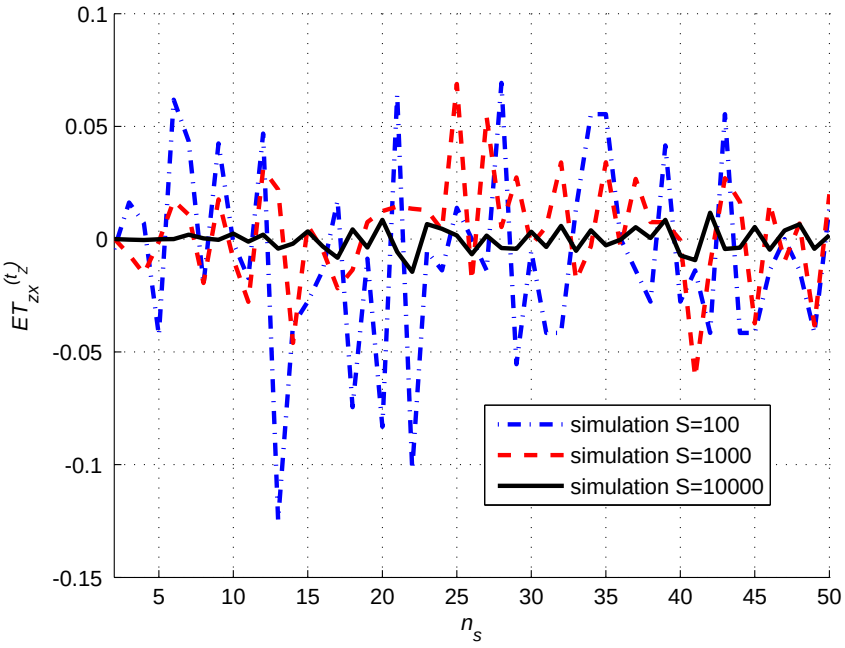

Figure 8.16: Effective Transfer Entropy $ET_{ZX}^{(t_Z)}$ versus number of states n_s for null model. Analytical values are 0 and simulated values acquired using equation (8.1) on simulated data of varying sample size S.

The corrected Transfer Entropy suggested in [82] generalizes the effective Transfer Entropy by taking the average values of M permutation surrogates instead of just one realisation such that

$$ET_{YX}^{(\tau)} = T_{YX}^{(\tau)} - \sum_{i=1}^{M} T_{Y_{S_i},X}^{(\tau)} \tag{8.2}$$

where Y_{S_i} is the ith randomly shuffled surrogate of time series Y. The reasoning behind this is that surrogate have bias of their own [101] and by taking the average of different realisations the bias and variance is reduced producing a much stable and smooth estimate of Transfer Entropy on the shuffled surrogate. From Figure (8.16), using sufficient surrogate estimate, it may be possible to identify 0 values of Transfer Entropy on the toy model where overestimation is only due to insufficient data to get good probabilities. However, in real data sets there are many other factors to be taken into account in terms of obtaining good probability.

8.4 Estimation of Entropy

The estimation of the information theoretic values on real data sets is where the true challenge lies. The finite sampling effect is just one of many problems faced in estimating probability of real data sets. In the toy model and the Ising model we knew exactly what the numbers of states n_s and what the discrete values of the states were. In EEG data sets, most of the values are approximately continuous and to account for each state separately would generally not be feasible. State space will need to be reconstructed with certain estimates. There exist a whole range of literature with regards to entropy estimates, a good summary of entropy estimation in relation to causality detection is given in [52].

Entropy is simply the expectation of log values of probabilities. Therefore entropy depends completely on probability estimation and so does the other entropy based measures such as Mutual Information and Transfer Entropy. Analytical values of Mutual Information and Transfer Entropy can be defined for discrete values as we have done in previous chapters and also for continuous values [57, 32, 33]. Parametric estimations works directly with continuous values when there are reasons to believe that assumption of a certain distribution may be true. The most common assumption is the Gaussian assumption [7]. The Edgeworth expansion is an example of a parametric estimator which approximates entropy through asymptotic expansions [95, 96]. Unfortunately on EEG data sets there is no reason to expect any type of underlying distribution and therefore we proceed with the nonparametric estimators. Non parametric estimators that will be mentioned here includes histogram, nearest neighbour estimates, rankings and kernel estimation methods.

8.4.1 Classical histogram (equidistant binning)

Probabilities of discrete values are relatively easy to obtain. Therefore coarse graining techniques converting continuous (or approximately continuous) data into discrete states are often utilized [57], so that the data can be treated as discrete values. This is done with the assumption that the coarse grained values converges to continuous values as the coarse graining gets more and more refined. For classical histogram, convergence of Mutual Information and Transfer Entropy estimates to continuous values have been theoretically proven [57] . The action of course graining is the partitioning of the continuous data

in order to use discrete probability estimation tools. This is also known as the state space reconstruction [79].

In the toy model, n_s was the number of states that we had on the model. In real data sets with continuous values, one can never have enough states to accommodate all the possible values hence coarse graining is applied. Our approach to coarse graining would be to set the values of n_s that is required and then divide the interval between the maximum and minimum amplitude into n_s equal sized bins where the values will be grouped accordingly. The probabilities will then be obtained by counting the visitation frequencies. This approach is known as the classical histogram or equidistant binning approach. It is said to be the simplest [101] and most widely used form of course graining [21]. For an interesting example see [74], where classical histogram is utilized for estimation of Mutual Information on EEG data sets.

For example if there are two processes X and Y, the probabilities will be obtained by counting the numbers of values in various bins that were obtained by partitioning the range of X and Y into finite size bins. Let $i, j \in \{1, \cdots n_s\}$ such that $C_X(i)$ be the number of values falling into the ith bin of X, $C_Y(j)$ is the number of values falling into the jth bin of Y and $C_{XY}(i, j)$ is the number of values in their intersection. Then the probabilities are approximately $p_X(i) \approx \frac{C_X(i)}{C}$, $p_Y(j) \approx \frac{C_Y(j)}{C}$ and $p_{XY}(i, j) \approx \frac{C_{XY}(i,j)}{\grave{C}}$ where C is total number of values and $\grave{C}$ is the total number of pairs. The Mutual Information estimation is then obtained by

$$I(X, Y) \approx I_{binned}(X, Y) = \sum_i \sum_j p_{XY}(i, j)\, log \frac{p_{XY}(i, j)}{p_X(i) p_Y(j)}. \tag{8.3}$$

Similarly, the Transfer Entropy estimation can also be obtained.

Having uniformly sized bins is the simplest implementation of this approach [21, 105], there are alternative ways of partitioning into unequal sized bins. This is known as adaptive binning, one example on Mutual Information have been proposed by [39], where boxes are subdivided only locally in places where the structure is statistically significant in order to avoid too few sample points in a certain bin. Other algorithm for adaptive binning are explored in [88, 30, 23, 101]. As n_s grows bigger and bigger, the actual bin size gets smaller and the values for the estimated measures should converge to the continuous values.

It has been theoretically proven in [57] that uniform partition converges for both Mutual Information and Transfer Entropy. However for adaptive partitioning convergence is only proven for Mutual Information but not for Transfer Entropy. This is one of the reasons we shall choose to use uniform partitions.

8.4.2 Rankings and symbolic analysis

Symbolic analysis is a special way of partitioning the state space. Firstly, coarse graining is done as usual. The difference is that these symbols are given ranks enabling the coarse grained data to be arranged in ascending or descending order. Each value is replaced by it's rank in the sorted sequence. The ranking converts any type of arbitrary probability distribution into uniform distribution [52].

A variable called permutation entropy [9, 48, 82] can be defined to measure the information on the order relations of the symbols. This variable is defined just like the Shannon entropy in equation (2.3a) except that it is the probability of the orderings between values that are taken into account instead of the probability of the values themselves. When permutation entropy is used instead of Shannon entropy in the definition of Mutual Information, conditional Mutual Information and Transfer Entropy, then the measures becomes Mutual Sorting Information, conditional Mutual Sorting Information [82] and symbolic Transfer Entropy [94] respectively.

In some special cases where duality between values and orderings can be established [48, 49] the symbolic Transfer Entropy is shown to be equal to Transfer Entropy. Indeed, the issues in Transfer Entropy estimations such as coarse graining, embedding vectors (the values to be conditioned on) and time delays persist for symbolic Transfer Entropy. Thus generally, whatever one does with Transfer Entropy can be done with symbolic Transfer Entropy. In certain circumstances where ordinal time series are available it is logical to apply symbolic Transfer Entropy. The resulting uniform distribution makes it easier to deal with as well. Several other variations of symbolic Transfer Entropy is proposed in [65, 82, 79, 72].

8.4.3 Other nonparametric estimations

Kernel density estimation (KDE) estimates the probability density using a kernel K which must be a normalized probability density function. If S is the number of samples of variables X, the approximate density function [81] is $p_X(x) \approx \frac{1}{S}\sum_{i=1}^{N} K(x - x_i, h)$ where x_i is the ith sample of X and h is the bandwidth (kernel width parameter). For example when the kernel is Gaussian

$$p_X(x) \approx \frac{1}{S}\sum_{i=1}^{N} K(x - x_i, h) = \frac{1}{S(2\pi h)^{d/2}}\sum_{i=1}^{N} exp\left(-\frac{\|x - x_i\|}{2h^2}\right) \tag{8.4}$$

with d being the dimension and h in the Gaussian case is simply the variance. Similarly, the estimations are done for Y and the joint probabilities before putting them into the Mutual Information formula. The correlation integrals [100] used in Schreiber's original paper [89] to estimate Transfer Entropy is also a type of kernel estimator.

The k Nearest Neighbour (kNN) estimation is an example of a metric method of estimation. The algorithm proposed in [60] uses distance defined by $\|z\|_z = max\{\|x\|, \|y\|\}$ for a point $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, where $\|.\|$ denotes Euclidean norm. Define $\mathcal{N}_k(i)$ to be the set of nearest neighbour samples of $z_i = (x_i, y_i)$ with respect to the norm $\|.\|_z$. Let

$$\epsilon_x(i) = max\{\|x_i - x_{\tilde{i}}\| \,|(x_{\tilde{i}}, y_{\tilde{i}}) \in \mathcal{N}_k(i)\}, \tag{8.5}$$

$$\epsilon_y(i) = max\{\|y_i - y_{\tilde{i}}\| \,|(x_{\tilde{i}}, y_{\tilde{i}}) \in \mathcal{N}_k(i)\} \tag{8.6}$$

so that we can calculate the number of elements within a distance for x and y

$$n_x(i) = |\{z_{\tilde{i}} \,|\, \|x_i - x_{\tilde{i}}\| \leq \epsilon_x(i)\}|\,, \tag{8.7}$$

$$n_y(i) = |\{z_{\tilde{i}} \,|\, \|y_i - y_{\tilde{i}}\| \leq \epsilon_y(i)\}| \tag{8.8}$$

and an estimator for Mutual Information formula in equation (2.12) is given by,

$$I^{(1)}(X,Y) \approx \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)) + \psi(n_y(i))], \tag{8.9}$$

$$I^{(2)}(X,Y) \approx \psi(k) + \psi(n) - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)+1) + \psi(n_y(i)+1)] \tag{8.10}$$

where ψ is the digamma function such that $\psi(x) = \Gamma(x)^{-1}d\Gamma(x)/dx$, $\psi(x+1) = \psi(x) + 1/x$ and $\psi(1) = -C$ where $C = 0.57772156$ is the Euler-Mascheroni constant. This algorithm has been generalized for conditional Mutual Information in [40]. Some advantages of the kNN estimator is that it has small bias for small k values and the fact that it is designed so that individual error of entropy estimations cancels out. It has also been reported that kNN performs better than the KDE given that k is appropriately chosen [95, 8, 40]. However, as far as our knowledge goes, there is no systematic strategy to choose k.

There are many other alternatives trying to address the various deficiencies in these estimation methods. An estimator that utilizes density ratio estimation with maximum likelihood method is presented in [95, 96]. Plug in estimates where consistent density estimations are substituted for actual densities are discussed in [18, 52]. No matter what the method there is always a parameter that needs to be decided upon, for example n_s for histograms, k for kNN and h for KDE. The choice of these parameters depends on the size of data samples and also what level of variance and bias that one aims to achieve.

When choosing the value of a parameter one needs to strike a balance between bias and variance. Recall that the bias of an estimator is the difference between estimators expectation value and it's theoretical value. Obviously the smaller the bias the better. On the other hand the variance which is the range of the expectation value is also needed. An estimator has to be flexible to fit the data well, hence the need for a certain variability. Balancing the bias and the variance is a delicate process in any estimation. Here, we depend on our knowledge from experiments on the models to help us through applications on real data sets.

8.4.4 Transfer Entropy estimators

No single estimator can claim to be the best as each has it own parameters and complications to take into account [52]. Therefore we have decided to utilize the simplest and most common estimator, the classical histogram. In estimating Transfer Entropy, we also need to consider the fact that Mutual Information and Transfer Entropy has been utilized in many different ways usually involving summing up values of different time lags [105, 78, 47], normalizing the values [52] and subtracting or dividing values of opposite direction to attain a directionality index [93]. Not to mention the ones with corrections as discussed in subsection (8.3.2). In Schreiber's original definition in equation (4.2) where $T_{Y\to X} = E\left[log \frac{P(X_{n+1}=x_{n+1}|Y_n^{(l)},X_n^{(k)})}{P(X_{n+1}=x_{n+1}|X_n^{(k)})} \right]$. Clearly, in order to apply $T_{Y\to X}$ to processes X and Y in this manner, one needs to determine the order of both the Markov processes such that l and k are obtained. These $Y_n^{(l)}$ and $X_n^{(k)}$ values are also known as the embedding dimension. Schreiber warned that having large embedding dimensions may lead to major inaccuracies. Therefore as in the toy model we shall apply the simplest form Transfer Entropy estimate as in equation (4.5) where conditioning is minimized and the objective is to detect the causal lag. The act of utilizing various different time lags in embedding dimension is sometimes referred to as horizons [31, 52].

Chapter Summary

We have seen from simulations of the toy model for higher number of n_s that insufficient sample size leads to spurious values. On the other hand if n_s is too small, we may loose some of the information. Not only is this true for Transfer Entropy but for Mutual Information and covariance as well. One of the aims of the toy model was to shed some light on how large the bin size should be in relation to sample size to avoid finite sampling effects. In order to determine the appropriate values of n_s for a given sample size S (data set lengths) we shall refer to Figures (8.10) and (8.11). Another option is to use corrections in the form of surrogates or effective Transfer Entropy. On real data sets there is much more to be taken into account in addition to sample sizes. There exist many forms of possible estimation methods with their own pros and cons. Despite the difficulties in estimations,

there is overwhelming interest in information theoretic causal related measures and while we are fully aware of the various estimation methods available, for the rest of the thesis we shall utilize the most straightforward classical histogram method as well as the simple form of Transfer Entropy in equation (4.5).

Chapter 9

Application to EEG Data Sets

We have electroencephalography (EEG) data sets from recordings on 10 healthy subjects that were asked to do nothing but close and open their eyes for a certain period of time. For every subject we get two sets of EEG data, one with their eyes closed (EC data) and another with their eyes opened (EO data). Eight electrodes at 250Hz (4 milliseconds sampling rate) were placed on the scalp of each individual which is numbered as in Table (9.1). The approximate brain function of different electrodes in relation to Table (9.1) are highlighted in Table (9.2).

The EC and EO data were recorded for approximately 120 seconds in 4 millisecond time interval thus giving us sample size of approximately $S = 30000$ data points. Therefore for each individual we shall have 8 time series representing 8 areas of the brain for both EC and EO, each of length approximately 30000. The data, recording machinery and pre-

CORTEX	ELECTRODE	SIDE OF THE BRAIN
Frontal	1	Frontal Right (RF)
	2	Frontal Left (LF)
Central	3	Parietal Right (RC)
	4	Parietal Left (LC)
Temporal	5	Temporal Right (RT)
	6	Temporal Left (LT)
Parietal	7	Parietal Right (RP)
	8	Parietal Left (LP)

Table 9.1: Numbering and labelling of the electrodes

CORTEX	APPROXIMATE FUNCTIONS
Frontal	attention, planning, working memory and inhibition
Central	controlling movements
Temporal	sounds, languages and multi sensory integration
Parietal	visual, spatial positioning and short term memory

Table 9.2: Cortices and its approximate brain functions

processing of the data was provided by the team led by Björn Crüts at Biometrisch Centrum (BMC) [1]. Björn's team have kindly shared their data sets and also gave very valuable advice regarding the interpretation of the data and the outcome of the analysis.

9.1 Visualizing the data

First and foremost we need to visualize the actual data to understand the nature of it. In Figures (9.1) and (9.2) the first 1000 data points of subject 1 where electrodes 1 and 2 (frontal cortices) and electrodes 7 and 8 (parietal cortices) are visualized for EC and EO cases respectively. One can see that the data is sinusoidal in nature and that the amplitude of the EC data is mainly larger than EO especially for the parietal cortices that is supposed to be processing the visuals. This coincides with our implementation of 'causality' on the models, where the causal link is imposed by imposing restrictions on certain variables such that it cannot change according to its internal dynamics as much. Therefore if the values of EC seems to be changing more rapidly and more regularly this could probably mean that it is less restricted than that of the EO where information needs to be exchanged and causal links are present.

According to Björn's team, the difference in amplitude is due to a well known fact in the neuroscience community that the Alpha band will dominate the parietal cortices whenever the eyes are closed. There are various opinions [84, 13] on what is the actual frequency of the Alpha band is, however here we choose to stick with the advised range of 8 to 12Hz. The Alpha band is mainly a sine like wave that can sometimes be easily detected by looking at the EEG data itself as evident in Figures (9.1) and (9.2). The Alpha band is not the only frequency band common in EEG data sets. The other bands that we will use to show differences in Transfer Entropy values are the Beta band (12 to 20Hz) and the fast

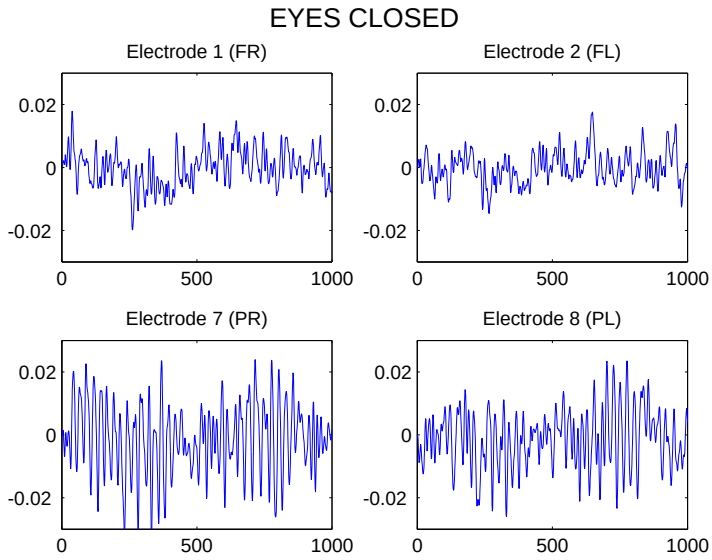

Figure 9.1: EC data of subject 1 from $\tau = 0 \cdots 1000$ for FR, FL, PR and PL

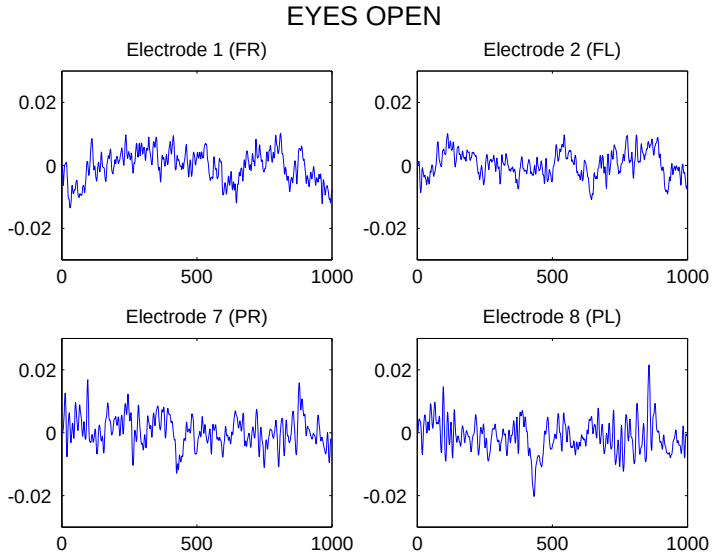

Figure 9.2: EO data of subject 1 from $\tau = 0 \cdots 1000$ for FR, FL, PR and PL

Beta band (20 to 32Hz).

In order to investigate the effects of these frequency bands, filters will have to be utilized. There are many different types of filters with their own strengths and weaknesses. Here we chose to use the Fast Fourier Transform (FFT) and the inverse Fast Fourier Trans-

form (iFFT) for all the frequency filtering mainly because we consider it to be the simplest form of filtering and it is readily available in MATLAB.

9.1.1 Transfer Entropy on sine waves

The sine-like pattern that appears in the EEG data for these electrodes will effect the Transfer Entropy estimations. One can see this by using Transfer Entropy between two sine functions. Replicating the data with a 250Hz sampling rate and $S = 30000$, let A be a sine function with frequency 10Hz and B be a sine function with frequency 4Hz. The first 100 time steps (4 milliseconds each) of A and B are visualized in Figure (9.3).

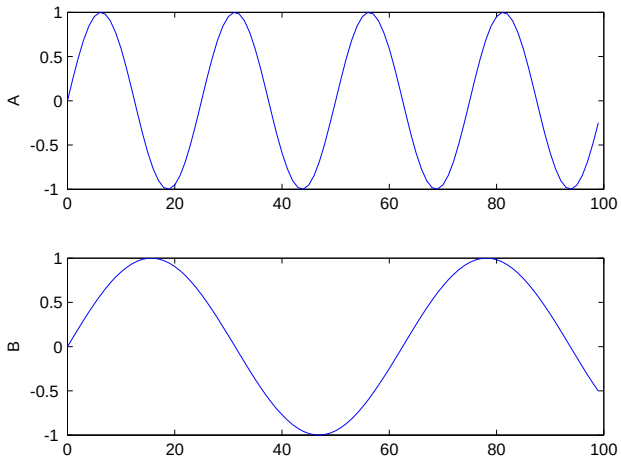

Figure 9.3: The first 100 data points of two sine waves with different frequencies

The Transfer Entropy values between A and B displayed in Figure (9.4) shows that the resulting Transfer Entropy estimations for both directions has 16 peaks as a result of adding the four peaks of A to the two peaks of B and multiplying the sum by two (account for both peaks and troughs since Transfer Entropy is positive definite). This is due to the fact the estimations of Transfer Entropy in (4.5) is supposed to detect patterns over certain time lags and the cyclical sine waves contributes to this (since we use the time average).

We then look at the estimations of $T_{AA}^{(\tau)}$ in Figure (9.5) which is supposed to be 0 by definition. It has roughly 8 repeated cycle of patterns due to the 4 peaks and 4 troughs in A. By using larger and larger n_s we expect the values to get values closer and closer

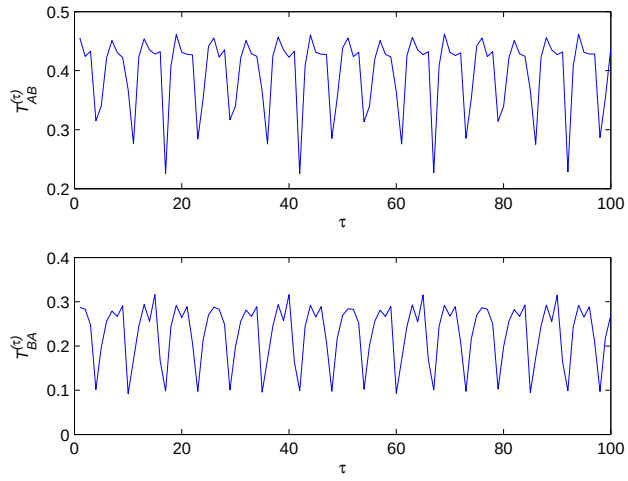

Figure 9.4: $T^{(\tau)}_{AB}$ and $T^{(\tau)}_{BA}$ using $n_s = 10$ to approximate data for $\tau = 0 \cdots 100$

to analytical value 0. At first glance this appears to be the case but this is not the case since estimations for $n_s = 25$ are larger than values for $n_s = 20$ when we expect it to be smaller. The estimated values when using $n_s = 30$ are smaller again. This confusing indication in terms of values of n_s to be used is most probably because of the finite sampling effect discussed on the toy model. One particular thing these values have in common is the minimum value of 0 at four points in the figure. Thus using corrections of shuffled surrogates will give negative values and could potentially be even more confusing.

We look once again at the estimations of $T^{(\tau)}_{BA}$, now with various n_s values in Figure (9.6) where do not expect the values to converge to 0. One can see that the estimations for $n_s = 20, 30$ leads to spurious values since the maximum value of Transfer Entropy is only $log(2) = 0.6931$ as discussed in the toy model and indicated in Figures (8.10) and (8.9). However, the larger n_s estimations does indicate that the values are converging towards a single value which is what we should expect.

So here is the conundrum, on one hand using larger n_s values lead to spurious estimations but smaller values have much higher variance. Moreover there is the question of how small a variance of the estimations on sinusoidal waves should be sufficient in order to be able to differentiate it from peaks due to causal lag. One important feature about all the estimations (even for smaller n_s values) is that the repeated patterns can be visible by using

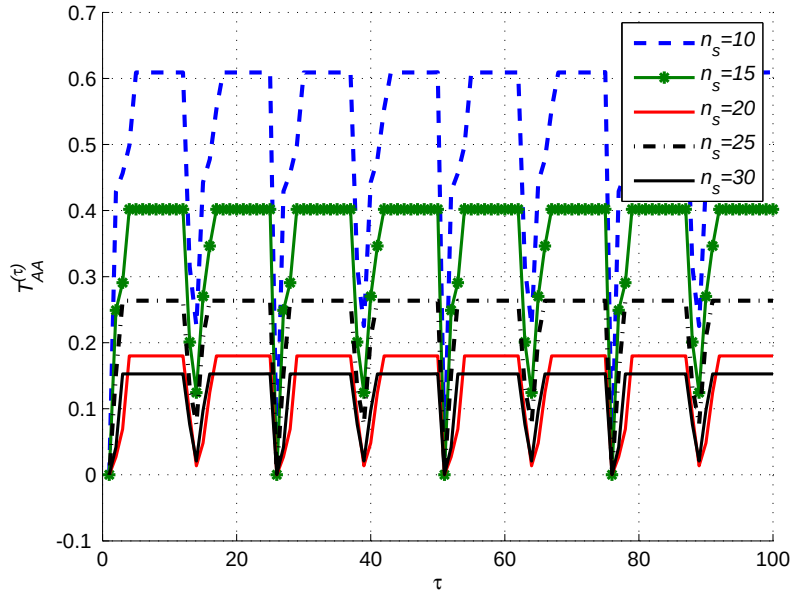

Figure 9.5: $T^{(\tau)}_{AA}$ using different number of states n_s to approximate data

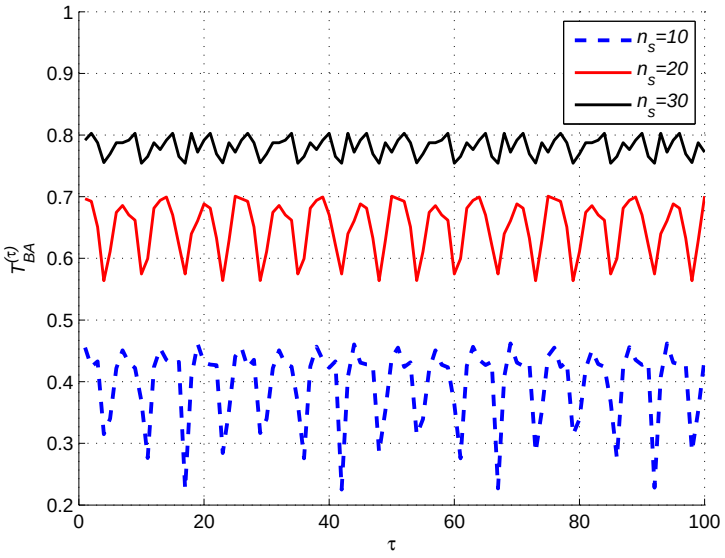

Figure 9.6: $T^{(\tau)}_{BA}$ using different number of states n_s to approximate data

$\tau = 100$ for large enough frequencies. Another feature of the Transfer Entropy of regular (unchanging height and phase like A and B) sinusoidal waves is that the amplitudes are more or less unchanged and changes could be indicated by damping effect.

9.1.2 Stationarity and ergodicity

As discussed in subsection (5.2.2), under certain conditions given by the ergodic theorem, the ensemble average is equal to the temporal average [77, 27]. On data sets, the ensemble average is obtained by averaging over different realisations of the data sets. The temporal average is where the probability of a variable is obtained by averaging the frequencies of different states over time. However, a prerequisite for the ergodic assumption is stationarity, and on EEG data sets stationarity is not always guaranteed. Statistical tests of stationarity has revealed a variety of estimates on the amount of time during which EEG remains stationary varying from several seconds to several minutes [14].

If one has enough realisations of a certain data set, ergodicity does not have to be assumed and the probabilities can obtained using the realisations (ensembles). However we only have data from 10 different subject. When different realisations are not available, what is usually done is that local stationarity is assumed for a certain range of time lags and then averages over moving windows will give meaningful result despite statistical errors [57]. Choosing the size of the moving window is also delicate process, since on one hand we have seen that insufficient statistics leading to spurious values and on the other hand we need the data to be local enough to capture the dynamics. Therefore if we were to use time windows some form of correction will need to be utilized. We have tried using moving windows of lengths 1000 to 5000 (with corrections) with results mimicking those obtained by using the whole length of data which is approximately 2 minutes. Due to these reasons we have decided to simply use to whole length of the data.

Taking lessons from the toy model, we proceed with the appropriate number of bins for the available sample size. If we were to use a simple classical histogram Transfer Entropy estimate on the whole length of data sets, from Figures (8.10) and (8.9), to avoid confusing finite sampling effects, it looks like $n_s = 10$ would be the safest value to choose. Moreover from the previous discussions of Transfer Entropy on sinusoidal waves, in certain cases one can distinguish which effects comes from the sinusoidal nature. Furthermore all Transfer Entropy results given in this chapter is averaged over the 10 subjects.

9.2 Transfer Entropy between hemispheres of the brain

Naturally we will first look at the parietal cortices (electrodes 7 and 8), the ones that are supposed to be processing the visuals. Afterwards we move on to the frontal ones (electrodes 1 and 2) which exhibits interesting differences between the EC and EO data.

9.2.1 Transfer Entropy of parietal cortices

We display values obtained by utilizing the Transfer Entropy formula in equation (4.5) for different frequency ranges in Figures (9.7) and (9.8). Firstly we do this on the raw data (incorporating all frequencies) for the graph labelled ALL. The graph labelled ALPHA is the Transfer Entropy on the data obtained by filtering out the other frequencies save the Alpha band between 8Hz and 12Hz. Similarly, the graph labelled BETA is Transfer Entropy when other frequencies are filtered out except the Beta band (between 12Hz and 20Hz) and the FASTBETA graph is Transfer Entropy when only the fast Beta band (between 20Hz and 32Hz) is preserved.

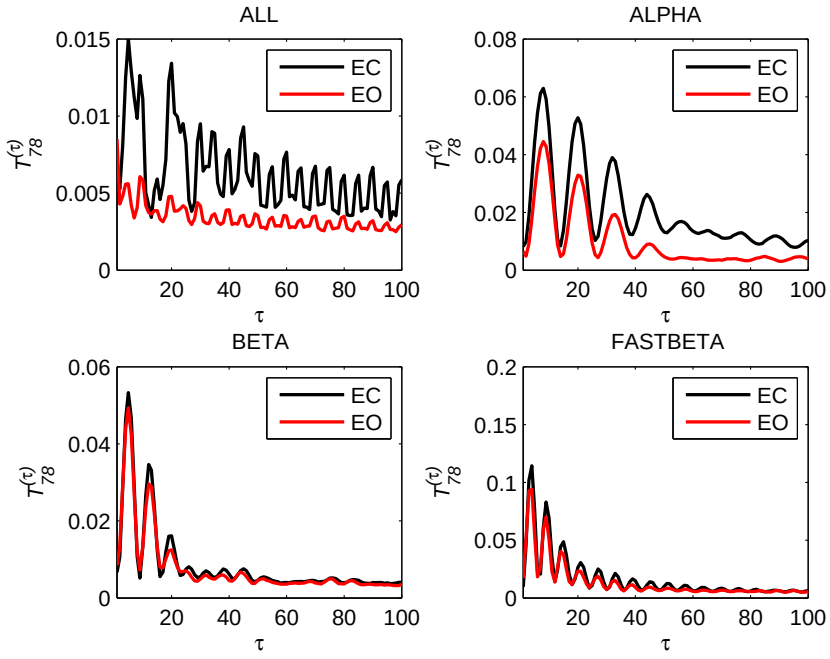

Figure 9.7: $T_{78}^{(\tau)}$ (PR $\rightarrow$ PL) for $\tau = 0 \cdots 100$ for different frequency ranges

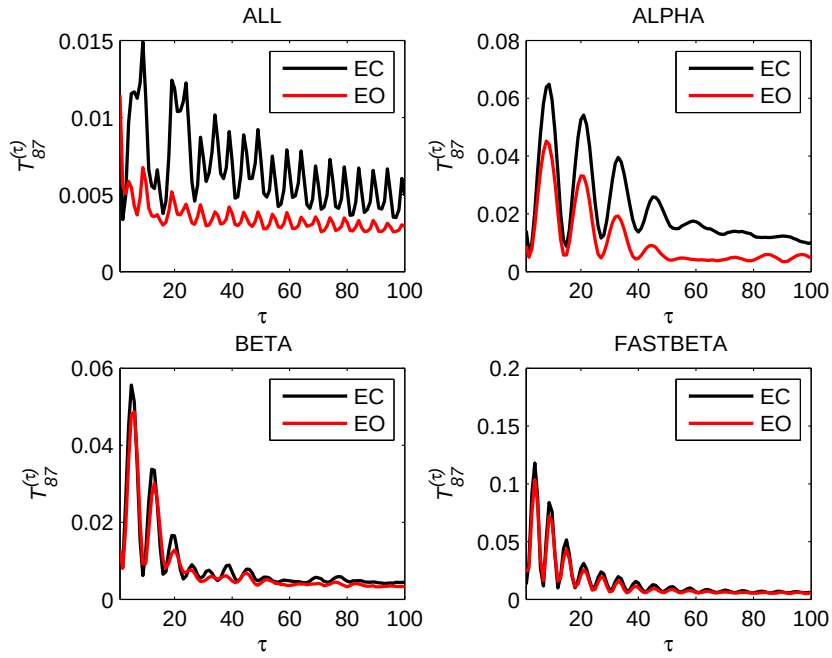

Figure 9.8: $T_{87}^{(\tau)}$ (PL $\rightarrow$ PR) for $\tau = 0 \cdots 100$ for different frequency ranges

What one can clearly see in Figure (9.7) and Figure (9.8) is that calculating Transfer Entropy within different frequencies render different results and that generally the Transfer Entropy values of EC is bigger than EO especially for the Alpha band where the relative difference is clear and the values are not converging even at $\tau = 100$. We suspect the main reason is because the frequencies are more regular in EC data. Another observation of Figure (9.7) and Figure (9.8) is that Transfer Entropy values in the ALL graph are not only very much influenced by the alpha band frequencies, but also by the electrical mains that will influence the data at 50Hz frequency therefore this needs to be filtered out. One common thing about all the graphs is that the sinusoidal values are damping out, possibly indicating that influences (amplitudes) of certain frequencies only last for a certain amount of time lag, approximately $\tau = 50$ i.e. around 0.2 seconds lag.

The question is now whether the ALL graph in Figure (9.7) and Figure (9.8) is simply a superposition of all the frequencies values or can some emergent causal behaviour be observed. To better see the underlying differences we shall venture to filter out some the frequencies. We apply upper bound of 32 Hz and lower bound of 1 Hz as well as filtering

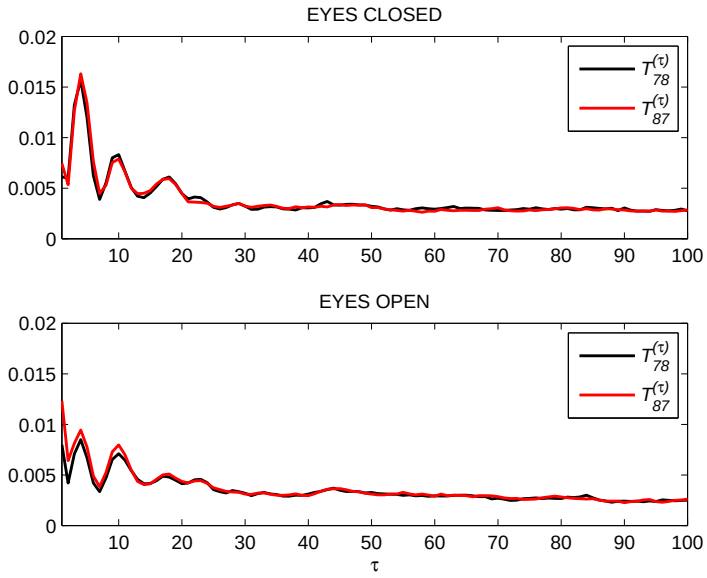

Figure 9.9: $T_{78}^{(\tau)}$ and $T_{87}^{(\tau)}$ for $\tau = 0 \cdots 100$ without the Alpha band

the Alpha band out to get Figure (9.9). Both directions of Transfer Entropy are displayed together in the figure and the values are almost identical for both direction where the amplitude seems to be dampening out. From this figure we may think that nothing much else happens in frequencies other than the Alpha band and perhaps causality between the two nodes of the parietal lobes really does depend only on the Alpha band. Before coming to any conclusion, we continue to look at some other electrodes.

9.2.2 Transfer Entropy of the frontal cortices

Now we focus on electrodes 1 and 2 that roughly represents the frontal cortex. Recall that the frontal cortices are supposed to be controlling attention, planning, working memory and inhibition. We begin by displaying Figure (9.10) and Figure (9.11) which is the counterparts of Figure (9.7) and Figure (9.8) on frontal cortices. However, a striking difference is that the ALL and ALPHA graphs in both Figures (9.10) and (9.11) looks more like a different approximation (using different amplitudes) for the same underlying values, whereas the Transfer Entropy values in the ALL and ALPHA graphs of EC and EO data in Figures (9.7) and (9.8) looks like they are distinctly approaching a different value. This is more

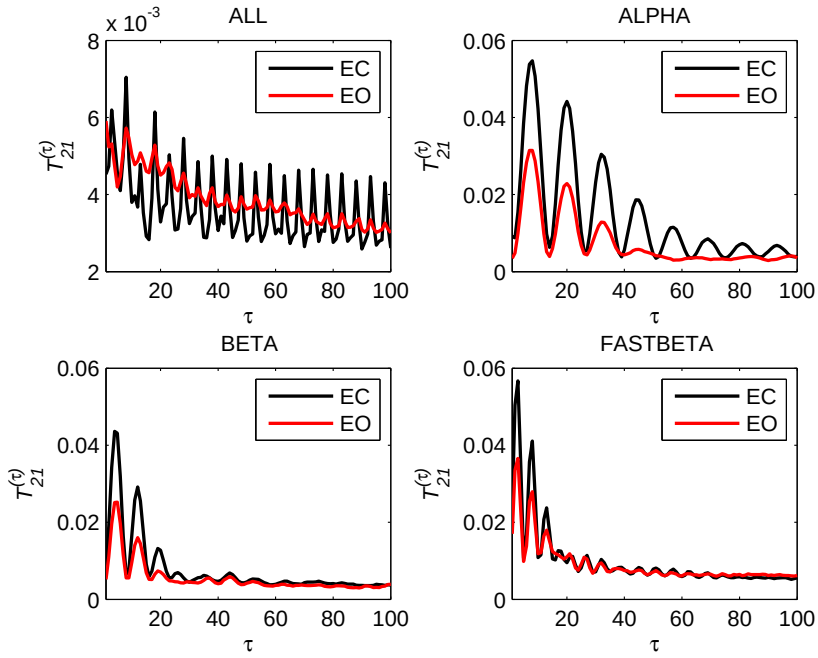

Figure 9.10: $T_{21}^{(\tau)}$ (FL $\rightarrow$ FR) for $\tau = 0 \cdots 100$ for different frequency ranges

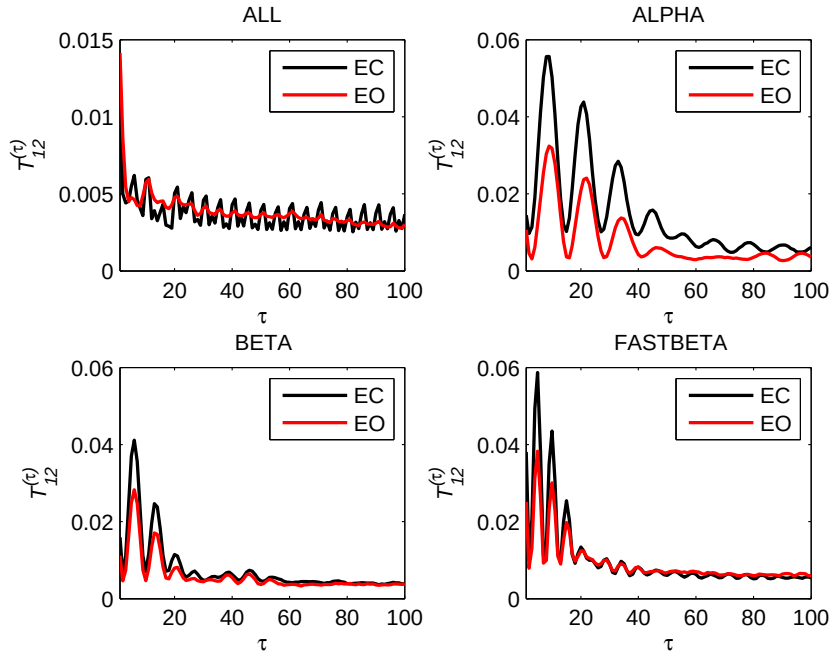

Figure 9.11: $T_{12}^{(\tau)}$ (FR $\rightarrow$ FL) for $\tau = 0 \cdots 100$ for different frequency ranges

obvious for the ALPHA graph in Figure (9.10) than for the one in Figure (9.11) and this could imply more FR $\rightarrow$ FL causation in the Alpha band for EC than EO. There does not look like there is much difference in terms of the other bands.

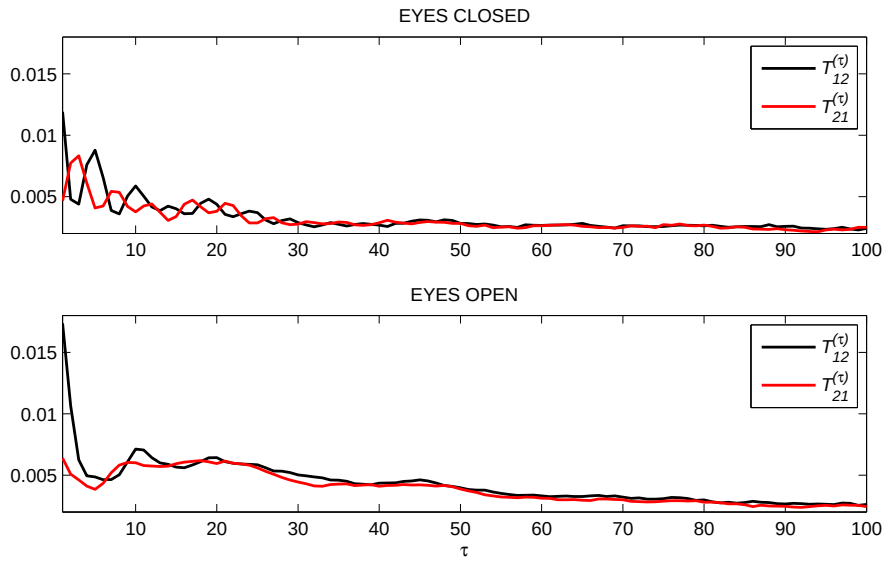

Figure 9.12: $T_{12}^{(\tau)}$ and $T_{21}^{(\tau)}$ for $\tau = 0 \cdots 100$ without Alpha band

In Figure (9.12), we display the values of Transfer Entropy for both directions on data filtered for and upper bound of 32Hz and a lower bound of 1Hz as well as the Alpha band between 8Hz and 12Hz. One can clearly see that the EO and EC is different even though the Alpha band is not there. In the EC data the initial Transfer Entropy entropy peaks seems to be alternating in their values but not so much in the EO data. There could be a few explanations as to the reason of this. One explanation is that the alternating values in EC is normal pattern of information exchange (or non exchange) between the two hemispheres which is disrupted in EO due to the need to process the information obtained. Another possible explanation for the alternating values is that in EC the two hemisphere are simply out of phase, in this case by approximation $\tau = 2$ (8 milliseconds). This could be due to different interaction rate with other electrodes. However if we look at Figures (9.10) and (9.11) we shall see that both EC and EO are out of phase. Therefore the difference between EO and EC in Figure (9.12) must be due to actual difference in the dynamics. We now proceed to look at interactions between parietal and frontal cortices.

9.2.3 The interaction between frontal and parietal cortices

We now focus on the interaction between the frontal and parietal cortices where we shall look at the right part of the brain (electrodes 1 and 7) and the left part (electrodes 2 and 8) separately. Figures (9.13) and (9.14) looks at Transfer Entropy values where the parietal cortices causes frontal ones. Once again the graph labelled ALL refers to the Transfer

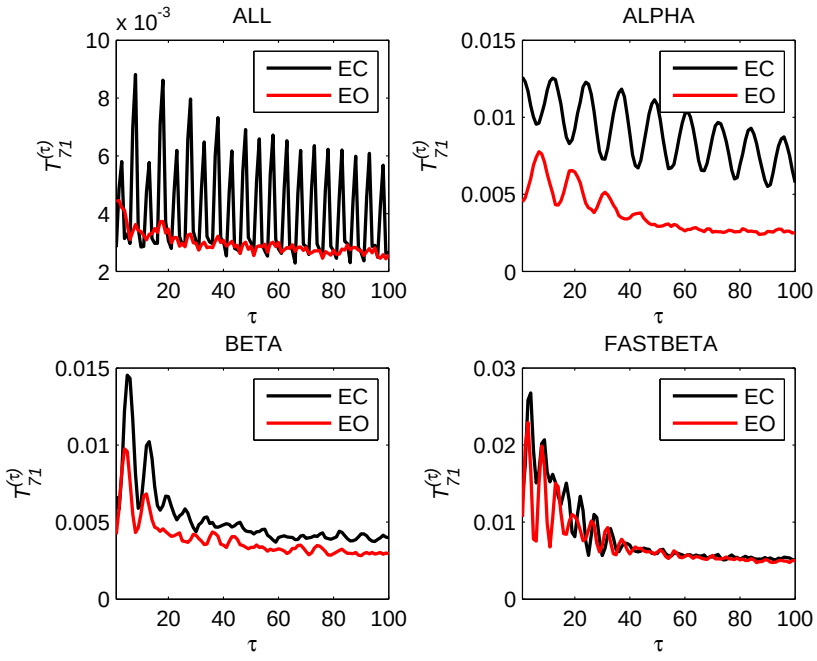

Figure 9.13: $T_{71}^{(\tau)}$ (PR $\rightarrow$ FR) for $\tau = 0 \cdots 100$ for different frequency ranges

Entropy applied to the whole range of frequencies and the graph labelled ALPHA is the Transfer Entropy on the data obtained by filtering out the other frequencies save the alpha ones between 8Hz and 12Hz. Similarly the graphs labelled BETA and FASTBETA were obtained by focusing on their respective frequencies. Figures (9.13) and (9.14) are different from the graphs of previous section because the former is exclusively on the right hand side of the brain and the latter is on the left side and not the interaction between the two sides. The first thing to notice is that the amplitudes of the values are all graphs in both figures are generally smaller than amplitudes in Figures (9.7), (9.8), (9.10) and (9.11). As we can see there are some differences between the two interactions however the similarities are more prominent especially in seeing that the Alpha band values are definitely larger for EC data

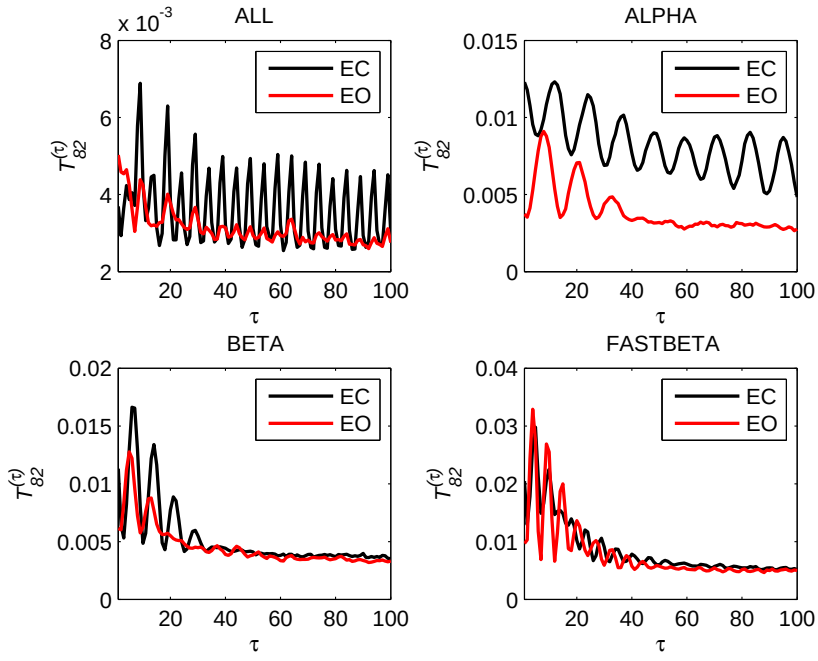

Figure 9.14: $T_{82}^{(\tau)}$ (PL $\rightarrow$ FL) for $\tau = 0 \cdots 100$ for different frequency ranges

and the sinusoidal effects lasting relatively much longer than the EO values.

The interaction between the parietal and frontal cortices when Alpha bands are filtered out is displayed in Figures (9.15) and (9.16). At first glance it show clear difference between EC and EO even with the absence of Alpha band. We have similar behaviour in alternating peaks in EC like in Figure (9.12) and the interpretation could be of similar nature as well. In the EO there is one clear peak and trough before a kind of lump in both Figures (9.15) and (9.16). This is one clear peak happens at $\tau = 3, 4, 5$ in the EO data of both figures before the lump between $\tau = 10$ and $\tau = 30$ (a time lag of 80 milliseconds) where values in both direction are more or less equal. Although directionality cannot be inferred since values in both direction is equally high, it is very possible that information is exchanged in both directions. One can speculate that since this lump comes after the initial peak and trough, it is the rapid exchange of information in deciphering visual data from the parietal cortices which disrupts the patterns of EC in relation to the idea that 'causality' is a form of restriction on changes of the affected variables.

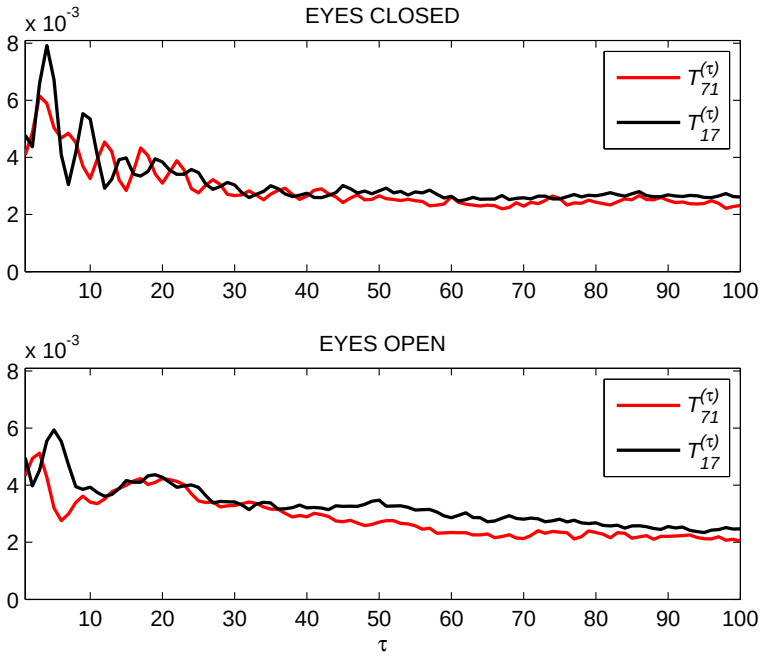

Figure 9.15: $T^{(\tau)}_{71}$ and $T^{(\tau)}_{17}$ for $\tau = 0 \cdots 100$ without Alpha band

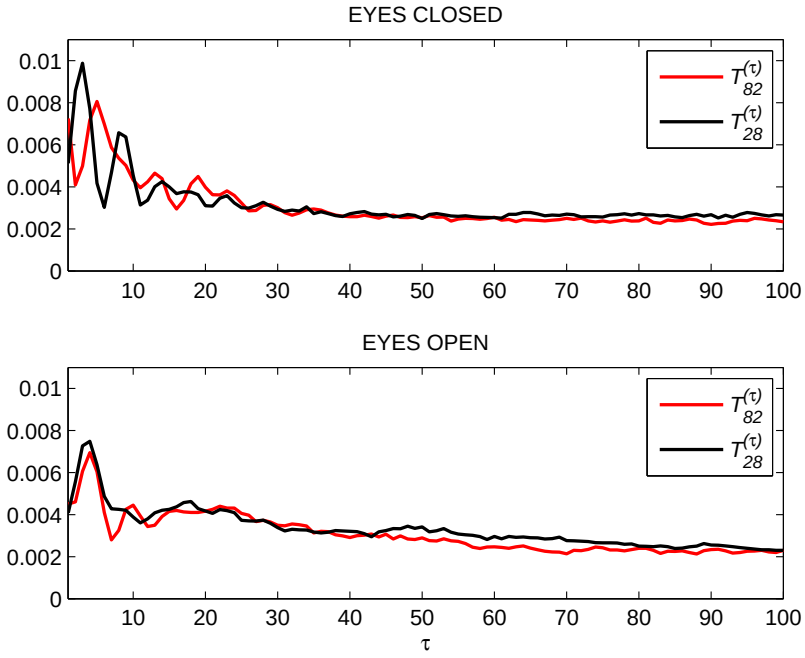

Figure 9.16: $T^{(\tau)}_{82}$ and $T^{(\tau)}_{28}$ for $\tau = 0 \cdots 100$ without Alpha band

9.3 Discussion

From Figures (9.7), (9.8), (9.10), (9.11), (9.13) and (9.14) one can see that Transfer Entropy on the data set are different when estimated within different frequencies. The obvious difference between EC and EO data lies in the Alpha band and this begs the question wether causality in this case depends solely on the Alpha band frequencies. It seems that this might be the case for the interactions between the parietal cortices. However, from Figures (9.12), (9.15) and (9.16) we say that the answer is no. Even without the Alpha band the difference between EC and EO data is clear. The difference is actually the clearest in the frontal cortices.

9.3.1 Frontal cortices

All three Figures (9.12), (9.15) and (9.16) is related to the frontal cortices. However, the most striking difference between EO and EC can be seen when the Transfer Entropy is used on itself.

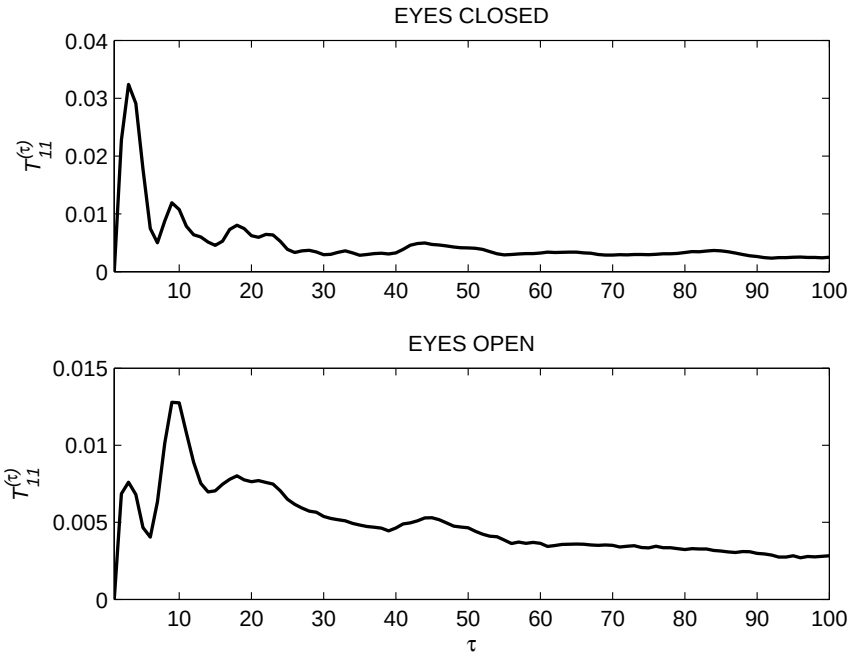

Figure 9.17: $T_{11}^{(\tau)}$ for $\tau = 0 \cdots 100$ without Alpha band

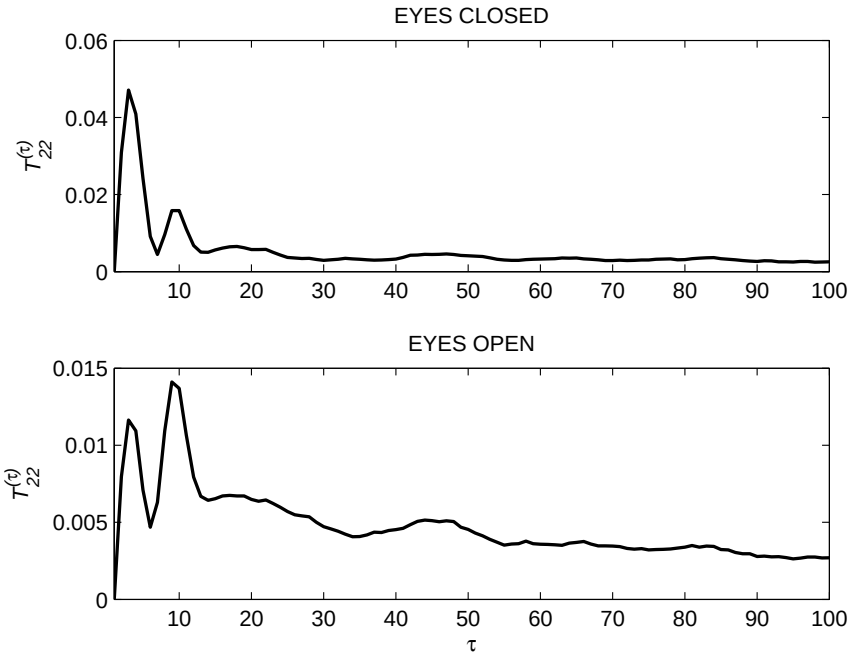

Figure 9.18: $T_{12}^{(\tau)}$ for $\tau = 0 \cdots 100$ without Alpha band

Figures (9.17) and (9.18) are the Transfer Entropy values on the data filtered for upper and lower bound as well as the Alpha band. To say that the peaks in the figures are simply a reflection of Transfer Entropy in the Beta and fast Beta bands is quite difficult in this case since we see that in Figures (9.19) and (9.20) that the Beta and fast Beta band in the EC and EO are not very different in their behaviour. In fact even the Alpha band in the figures looks like its approximating the same value. It seems like something different has emerged in the cumulative frequencies that cannot be detected by examining the bands separately.

9.3.2 Causal lag detection

In the EO graph of Figures (9.17) and (9.18) the Transfer Entropy values peak at approximately $\tau = 10$ (40 milliseconds). There is also the second highest peak which actually comes first at $\tau = 3$ (15 milliseconds). One can interpret Transfer Entropy value when used on itself in accordance with equation (4.5) as a kind of feedback loop where the changes in current values depend on the values at τ previous time steps. Thus in a way, we can say that for EO data of both electrodes 1 and 2, the feedback loop (causal lag on itself) is approxi-

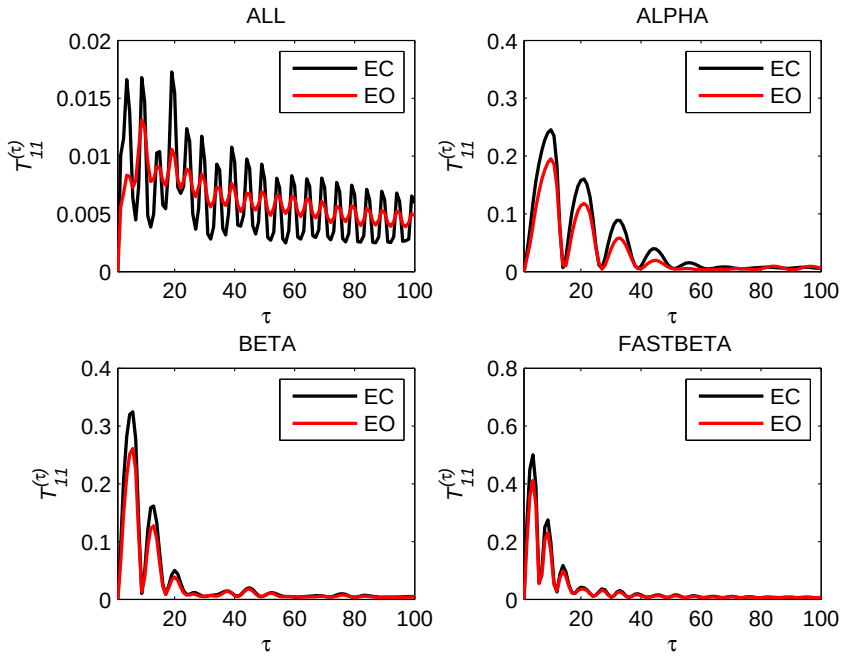

Figure 9.19: $T_{11}^{(\tau)}$ (FR $\rightarrow$ FR) for $\tau = 0 \cdots 100$ for different frequency ranges

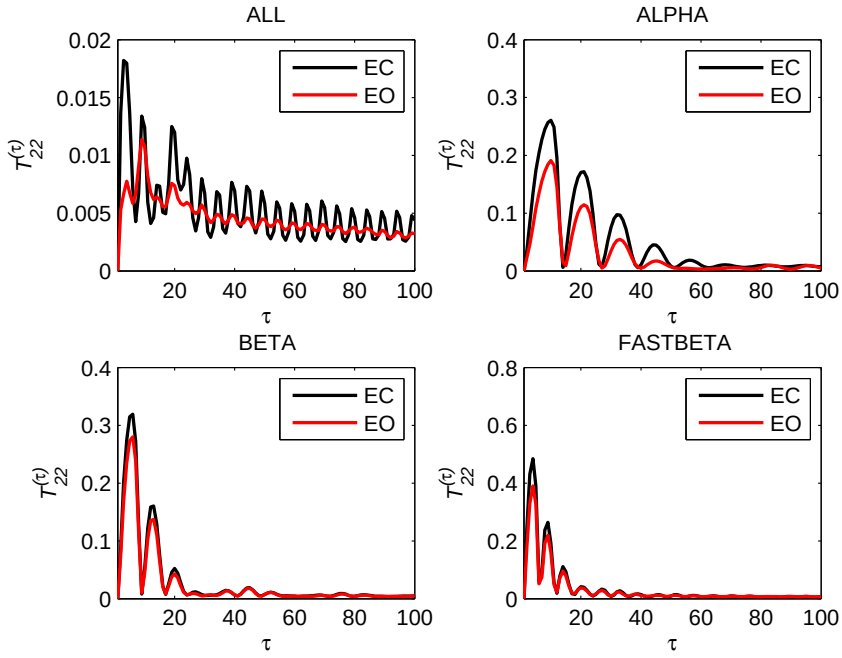

Figure 9.20: $T_{22}^{(\tau)}$ (FL $\rightarrow$ FL) for $\tau = 0 \cdots 100$ for different frequency ranges

mately 15 and 40 milliseconds. Why the second peak is higher than the first might have to do with the processing time of the information obtained from the parietal cortex. The good thing about these data sets is that the only difference between EO and EC is supposed to be in the visual and nothing else. Therefore based on that assumption, the changes in the frontal cortices should also be due to the extra information processing and communication due to visuals. $\tau = 10$ (40 milliseconds) could possibly just be the time the frontal cortices need to process a certain amount of visual before reacting (or causing itself to react).

We would like to think that this is also the case in for EO values Figures (9.15) and (9.16) with regards to the peaks at $\tau = 3, 4, 5$. It could possibly be the causal lags at which neuron of different parts of the brain communicate since 12 and 15 milliseconds could quite plausibly [12, 22, 67] be the cumulative neuron firing rates. However in this case the suspicion that this is simply the influence of other frequency bands is more probable since the the Transfer Entropy estimation of fast Beta band usually has peaks around $\tau = 4, 5$.

Chapter summary

We have seen that the Transfer Entropy values are different when different frequencies are filtered out. Although a lot EEG data analysis research focuses on the frequency domain, the outcome of our analysis points out that focusing on a single frequency band may not capture the bigger picture. The Alpha band seems to be very important in the parietal cortices and its interactions with the frontal cortices, however it does not seem to be very important in the frontal cortices which seem to be doing the processing. We conclude that Transfer Entropy on the combined frequencies gives more than Transfer Entropy on individual frequency bands. Furthermore we have identified a possible causal lag (feedback time) of the frontal cortices of 12 milliseconds and a possible processing time of 40 milliseconds as well as possible causal lag between frontal and parietal of 12 to 15 milliseconds with a possible processing time of 80 milliseconds.

Chapter 10

Conclusion and Future Research

In Chapter 1 to 3 we discussed issues of nonlinearity as well as ‘causality’ as we envisioned it to be in the brain and this led to the examining of Transfer Entropy in Chapter 4. We tested this measure on the Ising model in Chapter 5 as well as 6 and on the a toy model with analytical values in Chapter 7 and estimations in Chapter 8. In Chapter 9 we applied Transfer Entropy on EEG data sets with interesting results. There is much more to be done in furthering our understanding of Transfer Entropy and ‘causality’ in general.

Throughout the thesis we have set to define causality. Firstly as a sort of independence across time lags and then in relation to G-causality and Transfer Entropy as something that will affect prediction. By focusing on Transfer Entropy and transition probabilities we see that causality in this sense has a lot to do with changes in the affected variable that is caused by certain values of the causal variable. When it comes to translating this to be replicated on a model, we made certain values of the causal variable restrict the changes in the affected variable. It is crucial that both these variables maintain some stochastic element as a deterministic relationship cannot be considered causal. When looking at the data set which should differ only by one action of processing visual, we found out that when there is no visual input (eyes closed) the amplitude of the EEG waves are bigger and more regular. Therefore this coincides with our idea of causality being imposed by restrictions and where no information is received (therefore no causal link required), the behaviour of the data is unrestricted.

10.1 EEG data analysis

There are a few main issues that was of concern when analysing the data, firstly was of course estimation and how the sinusoidal nature of the data was affecting the estimation. The Transfer Entropy on sinusoidal values are interesting in itself since sine waves are actually deterministic and we have seen that deterministic values either gives 0 or constantly gives the maximum values as seen in Figure (7.6) on the toy model.

The difficulty of Transfer Entropy estimation on the sine function is in relation to determining the number of states n_s where variance of Transfer Entropy is small enough and making sure to exclude spurious effects (corrections may results in negative values). We saw from Figure (9.5) that this can be very tricky and to make any judgement based on Transfer Entropy amplitude in determining directionality at this stage could be misleading. This is due to the fact that the values of Transfer Entropy that should peak for the causal lag as in the models interfere with the sinusoidal nature of the data that effects the estimation of data done using time average. How small should an acceptable variance of Transfer Entropy be in order for the sinusoidal effect to be accurately differentiated from and causal lag remains the main question.

In fact, there is the question of whether there actually exist differentiable peaks at all. It is entirely possible that these part of the brain are in communication all the time and that it will be hard to detect any outstanding lag in Transfer Entropy values unless something very different happens (epileptic seizures for example). This is where complete understanding of the magnitudes of Transfer Entropy and the relative differences in values would come in very handy. Therefore a thorough investigation on the toy model in terms of understanding the values of Transfer Entropy and more detail testing on the data set is needed.

10.1.1 What is causality of EEG data sets?

More importantly, there is the question of what kind of causality exist in EEG data sets. Although we have assumed that the waves behaves in tandem with our definition of causality in relation to restrictions, there are also other possibilities. Is it possible that causality can exist in terms of communicating the lack of input through regular waves? Could causality occur both ways simultaneously?

We found out that there are certain frequencies that dominates for certain areas of the brain for certain task. The prime example here is the situation where the Alpha band is known to dominate the waves when eyes of the subject is closed and this is known to be clearest in the parietal cortices. We have seen that this influences most of the electrodes in the EC data. The question is, would this be considered a sort of causal influence of the parietal cortices (where Alpha band is strongest) to the other electrodes, after all it could still be signalling the other parts of the brain that the eyes are closed and are still closed over a period of time. Moreover some researchers that believe phase synchronization of two electrodes is sufficient to infer interaction [13].

If this is true then Figures (9.12), (9.15) and (9.16) would imply non interaction in most of the EC data since the phase is shifted by $\tau = 2$ (8 milliseconds) when both directions are compared. Or perhaps it could mean that information takes 8 milliseconds to go back and forth hence the shift. Moreover, in the EO data of Figures (9.12), (9.15) and (9.16), one can see that the initial peak and the lump that comes afterwards are consistent features for EO data of these three figures. This could imply communication in both direction is actively happening simultaneously. So now the question is, can causality occur in both directions at the same time? According to Schreiber's definition of Transfer Entropy in [89], directionality can only be concluded if one direction is determined to have 0 values and the nonzero in the other. Therefore according to him, in Transfer Entropy directionality can only occur is one direction. Again if one had a better understanding of the different magnitudes of Transfer Entropy perhaps different levels of causality could be determined. Hopefully with performing more experiments on different types of EEG data sets some of our suspicions could be verified or negated.

Moreover we have concluded that more than a single frequency band is needed in deciphering the causality for EEG data sets. It is our opinion that if estimated properly the Transfer Entropy on the accumulation of the various frequencies in the unfiltered data could be something very different than looking at the Transfer Entropy values of the frequency bands separately. It would be very interesting to see if the same results will be achieved if data sets with more complicated task should be tested. If the same conclusion can be achieved, this could imply that causality is an emergent property.

10.2 The models and its potential

The many factor influencing the magnitude of Transfer Entropy can include the individual dynamics of the causal and affected variable as well as the strength of the causal link (in the model indicated by how much restriction is imposed). For example in Figure (7.3) representing equation (7.1) we see that even the $T_{ZX}^{(\tau)}$ on the simple model depends on μ_X. Which means that the Transfer Entropy depends on the intrinsic probabilities (which is caused by individual dynamics) on the affected variable. In real data sets, this is something that we cannot measure unless we can find a way to distinguish individual dynamics from external influences. However from analysis on the data set we have seen that $T_{XX}^{(\tau)}$ is able to provide some interesting results. Therefore it would be very interesting to investigate more about the value of $T_{XX}^{(\tau)}$ and how it changes in relation to $T_{XZ}^{(\tau)}$ and $T_{ZX}^{(\tau)}$ as well as μ_X.

In the current toy model, we fixed the level of restriction that a certain stochastic process Z has over X and Y by controlling Ω which represents the percentage of Z values that allow changes in X and Y values. Therefore, $1-\Omega$ represents the percentage of Z values that restricts the changes in X and Y values. It is this restriction that creates the causal effect in the model. If we could somehow identify if there exist this form of restriction in the real data set then Ω could be used to gage the level of restriction and thus potentially the level of causality on real data sets. This could be very useful in addressing the problem related to the existence of causality in both directions and possible differentiation of weak and strong causal couplings.

Another finding on the toy model that can be proven analytically and alludes to the need of a formalism to quantify different levels of causality (or restriction) at least in terms of the amplitudes of Transfer Entropy, is the fact that there exist $T_{ZX}^{(\tau)}$ such that $T_{ZX}^{(\tau)} \neq 0$ even when $\tau \neq t_Z$. In the toy model this happens when $\mu_X = \frac{n_s-1}{n_s}$ and the influence comes through the variable $Q_{sgn(\gamma)}^{(\tau)}$ which represents the probability of the condition being fulfilled given the current information available time τ. Since $T_{ZX}^{(\tau)} \neq 0$ and $T_{XZ}^{(\tau)} = 0$, based on Schreiber's [89, 57] way of determining direction, one will have to conclude that there exist a direction for all these values of τ and not only at the causal lag t_Z. But the magnitude of Transfer Entropy correctly indicates that the direction (hence the causal link) is the strongest at t_Z and the other are just the side effects. This highlights the importance of

actually detecting causal lags in order to be able to distinguish side effects. We believe that further investigations in relation to the magnitudes of Transfer Entropy will be beneficial to the general formalism of quantifying causality.

The same can be said about the amended Ising model which provide a more realistic interactions in terms of the imposed 'causality' versus nearest neighbour interactions. We could try putting sites A, B and G further apart in a larger lattice or add much more interactions and link just to see the effect at T_c. Looking at the temperatures very close to T_c and working out an exponent for Transfer Entropy would also be an interesting direction to pursue.

10.2.1 Linking the model and data sets

On the models we have shown that the amplitude of Transfer Entropy can be affected by many factors which are very complicated to identify in the real data sets. One of the reason is that in the toy model we have set the distributions to be uniform so that values are simpler to estimate and the causal lag detection is more straightforward. The symbolic analysis and ranking discussed in subsection (8.4.2) is reported to convert any type of arbitrary probability distribution into uniform distribution [52]. If this is true then this symbolic Transfer Entropy could make our toy model even more relevant for real data applications. More importantly it might be able to provide us with some insight to Ω values defined on the toy model since we know that $\mu_X = \frac{n_s-1}{n_s}$ leads to uniform distribution for any variable X. Moreover, recently there have been a lot of interest in using the symbolic Transfer Entropy due to supposedly better estimations [65, 82, 79, 72] especially in the case of multi-fractal phenomena [71]. Therefore the symbolic Transfer Entropy certainly will be worth exploring in relation to the toy model.

Even if these values cannot be explicitly identified on the data sets, the toy model is no less valuable. What we have currently done is place the causality in one clear direction at one clear causal lag. However, in the brain it would be more logical to put more causal direction between many different stochastic processes and see how the direction and the influence clashes. The modelling possibility is endless, we could include different causal connections at different level of influences as determined by Ω and we could test the effects

of the stochastic processes being sinusoidal with dampening effects in order to gage the actual appropriate size of n_s needed in order to clearly establish causal lags. We think that the model itself can be a powerful tool in terms of aiding our understanding of causality not just in terms of Transfer Entropy but also in terms of general causality replication where any causal connections needs to be replicated in models.

10.3 Information theoretic measures

We have identified a measure that can capture features of the brain being nonlinear and 'causal' namely Transfer Entropy. There is much to be done, applying the measure with the knowledge obtained from the models remains one of the most important lesson of the thesis as different application may lead to different interpretation. Nevertheless the theoretical side remains as interesting as ever, with a lot more variations of Transfer Entropy definitions to test out and even more generalizations of Mutual Information are being proposed to unify frameworks of our understanding.

10.3.1 Variations of Transfer Entropy

There is so much more to be achieved by using these information theoretic measures in any type of data especially in neuroscience as demonstrated by the ample interest shown in the literature. Here, we propose some possible future directions that might be promising and interesting to pursue.

In this thesis one of our aim was to show that $T_{YX}^{(\tau)}$ will be largest at exact causal lag τ given that the change in X occurs directly at the next time step. We utilized the previous formula of Transfer Entropy as $T_{YX}^{(\tau)} = I(X, Y^{-\tau}|X^{-1})$ in equation (4.6), fully aware that this is not the only possibility of utilization. One interesting example would be, also varying the time steps for X such that

$$T_{Y\to X}^{(\tau_1,\tau_2)} = H(X|X^{-\tau_1}) - H(X|Y^{-\tau_2}, X^{-\tau_1}) = I(X, Y^{-\tau_2}|X^{-\tau_1}). \tag{10.1}$$

Another variation could be conditioning on two different variables

$$T^{(\tau_1,\tau_2)}_{Y,Z\to X} = E\left[log\ \frac{P(X_n = x_n|X_{n-1} = x_{n-1}, Z_{n-\tau_1} = z_{n-\tau_1}, Y_{n-\tau_2} = y_{n-\tau_2})}{P(X_n = x_n|X_{n-1} = x_{n-1})}\right] \quad (10.2)$$
$$= H(X|X^{-1}) - H(X|X^{-1}, Y^{-\tau_2}, Z^{-\tau_1}).$$

All this will have to be done is a systematic way so that differences and similarities in conjunction with equation (4.5) be fully understood.

10.3.2 Generalized Mutual Information

As we have mentioned before there are various generalizations of Mutual Information that are proposed within different envisioned unifying frameworks. These generalizations could also be tested on the Ising model and the toy model. There any many forms of attempted generalizations of Mutual Information. We discuss one example here.

Recall that $I(X,Y|Z) = E\left[log\frac{P(X,Y|Z)}{P(Y|Z)P(X|Z)}\right]$ for random variables X, Y and Z. It would be very interesting to compare this quantity to the actual Mutual Information and one way to do this is to define the Mutual Information of three variables. The Mutual Information of X, Y and Z can be defined [29] as

$$\begin{aligned} I(X,Y,Z) &= I(X,Y) - I(X,Y|Z) \\ &= E\left[log\frac{P(X,Y)}{P(X)P(Y)}\right] - E\left[log\ \frac{P(X,Y|Z)}{P(Y|Z)P(X|Z)}\right] \\ &= E\left[log\frac{P(X,Y)}{P(X,Y|Z)}\right] - E\left[log\frac{P(X)}{P(X|Z)}\right] - E\left[log\frac{P(Y)}{P(Y|Z)}\right] \\ &= E\left[log\frac{P(X|Z)}{P(X)}\right] + E\left[log\frac{P(Y|Z)}{P(Y)}\right] - E\left[log\frac{P(X,Y|Z)}{P(X,Y)}\right]. \end{aligned} \quad (10.3)$$

The equation captures how the probabilities are different when it is conditioned on Z. Note that this quantity is symmetric with respect to X, Y and Z since we have that

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

Figure (10.1) clearly depicts this in a set-theoretic setting. Another example of generalized

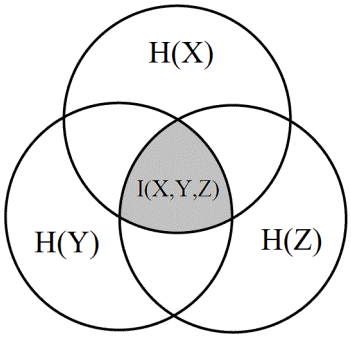

Figure 10.1: $I(X, Y, Z)$, $H(X)$, $H(Y)$ and $H(Z)$

Mutual Information has been introduced in subsection (2.4.2) in relation to clustering and another version known as Multi-Information is introduced by Ay in [35].

Conclusion

In Chapter 1, we outlined our view that even the most simplistic model of the brain should be nonlinear and should very logically be causal. Proceeding with the first assumption, we need a measure that captures nonlinearity hence Mutual Information whose variants include conditional Mutual Information and its time varied counterpart Transfer Entropy, was taken into account in Chapter 2. The second assumption of causality needs more expounding which led us to define causality very carefully in Chapter 3 as a dependency at a certain causal lag as proposed by Wiener and Granger ala G-causality. As opposed to G-causality which is linear by default, Transfer Entropy which can be interpreted as a nonlinear extension to G-causality seemed poised to capture both nonlinearity and causality.

Therefore in Chapter 4, we looked very carefully at this propounded Transfer Entropy and it's pitfalls. We decided to avoid some of the pitfalls by taking the simplest case and utilizing it for causal lag detection, bearing in mind that there are many other possible way of utilizing Transfer Entropy. This simple definition of Transfer Entropy was utilized for causal lag detection on the Ising model in Chapter 5 and to our knowledge we are the first to apply it in this manner. What we saw was that the Transfer Entropy and conditional Mutual Information gave identical indications and thus concluded that some element of time in the sense of induced causal lag needed to be introduced to the model. The amended Ising

model was produced in Chapter 6 by restricting two of the sites on the lattice. The two sites were made dependent on a third sites in order to create a 'causal' relationship. Transfer Entropy had no difficulty in detecting this form of 'causality' and our suspicions were verified. The Ising model being binary in nature means that we loose a lot of nonlinearity and $I = 0 \iff \Gamma = 0$. Thus, we need to be able to increase the number of states in order to further investigate the effects of nonlinearity.

The toy model in Chapter 7 allows us to do just this. Taking inspiration from our results on the Ising model, we set to model only three stochastic variables where two of them depend in a 'causal' way on one of the variable as in the amended Ising model. The toy model with two states ($n_s = 2$) or the simple model can be interpreted as the amended Ising model at higher temperatures where probabilities are uniform. If Z is the causal variable and X and Y are the affected variable then Ω stands for the percentage of states of Z that allows changes in X and Y and serves as an indication of the level of restriction imposed on the model at a chosen causal lag t_Z. It is this restriction that makes the relationship causal from a Transfer Entropy point of view.

In addition to that, $Q^{(\tau)}_{sgn(\gamma)}$ which represents the probability that the there are no restrictions on X and Y given the current knowledge available about Z at time lag τ, enables us to understand how the μ_Z influences $T^{(\tau)}_{ZX}$ when $\tau \neq t_Z$. We showed that through values of $Q^{(\tau)}_{sgn(\gamma)}$, $T^{(\tau)}_{ZX} \neq 0$ for values of τ close to t_Z. On the other hand, we also found out that at t_Z, μ_Z does not influence the value of $T^{(t_Z)}_{ZX}$ at all. More importantly, we were able to show that given that the causal lag t_Z, $T^{(t_Z)}_{ZX} \geq T^{(\tau)}_{ZX}, \forall \tau$ and therefore Transfer Entropy can be used for causal lag detection. This toy model can be modelled exactly and thus the simulation can be compared to its theoretical value. The simulations in Chapter 8 verified our theoretical formulation in Chapter 7, however it brings to light a problem in the form of finite size effects. As n_s gets bigger, more and more data is needed in order to obtain accurate probabilities and when the sample sizes becomes insufficient we get spurious values. One way to rectify this is by using surrogates.

On real data sets, there is no way to tell if exact probabilities are obtained. There exist many forms of possible estimation methods with their own pros and cons. We discuss some of the methods used for estimation of entropy in Chapter 8 and how the most common one is the classical histogram method. For data sets with continuous values, one does not

know what values of n_s to utilize on the data and this is where the null model is needed to provide information about the relationship between n_s and sample size. We decided to simply utilize $n_s = 10$ to avoid spurious values in applying the classical histogram method on the two data sets of EEG recordings from 10 subjects, one with eyes closed (EC) and the other with eyes opened (EO) in Chapter 9. While acknowledging that there is a lot of issues to be resolved in terms of estimation and what can be considered causality on EEG data sets, we point out that we have done a simple analysis that highlights the difference between the Transfer Entropy values of EO and EC on different frequency bands.

We observed that Transfer Entropy values are different for different frequency ranges and there are some frequencies bands that show more effect than the others and one obvious example is the Alpha band. Therefore we wondered if causality could simply be determined by a single band. However when examining the other frequencies we found out that even when the Alpha band is filtered out, the Transfer Entropy on the rest of the frequencies still indicates the differences between EO and EC. Thus we concluded that in this case causality is not something that can be determined by a single frequency band. In addition to that we have identified a possible causal lag (or feedback time) of the frontal cortices and its interaction with parietal cortices on the EO data which is in the order of 10 milliseconds.

We conclude that causality as determined by Transfer Entropy on EEG data sets is something very promising yet much more work should be done before anything strongly conclusive can be claimed. It would particularly interesting if causality can be viewed as an emergent property of different frequency bands and this is something that would probably be much clearer in EEG data sets with much more complicated task. The toy model developed in this thesis should be very helpful in furthering our investigations in doing a systematic analysis on the effects of Transfer Entropy in relation to ‘causality’.

References

[1] Biometrisch centrum: Hersenmetingen & neurofeedback. http://www.biometrischcentrum.nl. Accessed: 19/02/2013.

[2] D. Abásolo et al. Approximate entropy and auto mutual information analysis of the electroencephalogram in Alzheimers disease patients. *Med. Biol. Eng. Comput.*, 46:1019–1028, 2008.

[3] K. Anand and G. Bianconi. Entropy measures for networks: Towards an information theory of complex topologies. *Phys. Rev. E*, 80:045102, 2009.

[4] L. Angelini et al. Clustering data by inhomogeneous chaotic map lattices. *Phys. Rev. Lett.*, 85(3), 2000.

[5] L. A. Baccalá and K. Sameshima. Overcoming the limitations of correlation analysis for many simultaneously processed neural structures. *Prog. Brain. Res.*, 130:33–47, 2001.

[6] P. Bak. *How Nature Works: The Science of Self Organized Criticality*. Springer-Verlag, New York, 1996.

[7] C. R. Baker. Mutual information for Gaussian processes. *SIAM J Appl. Math.*, 19(2):451–457, 1970.

[8] R. C. Ball, M. Diakonova, and R. S. Mackay. Quantifying emergence in terms of persistent mutual information. *Advances in Complex Systems*, 13(3):327–338, 2010.

[9] C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.*, 88:174102, 2002.

[10] A. B. Barrett, L. Barnett, and A. K. Seth. Multivariate granger causality and generalized variance. *Phys. Rev. E*, 81:041907, 2010.

[11] A. B. Barrett and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. E*, 2009.

[12] J. M. Beggs and D. Plentz. Neuronal avalanches are diverse and precise activity patterns that are stable for many hours in cortical slice cultures. *The Journal of Neuroscience*, 24:5216–5229, 2004.

[13] J. Bhattacharya and P. Hellmuth. Enhanced phase synchrony in the electroencephalograph γ band for musicians while listening to music. *Phys. Rev. E*, 64:012902, 2001.

[14] S. Blanco. Time-frequency analysis of electroencephalogram series. *Phys. Rev. E*, 51(3):2624–2631, 1995.

[15] K. J. Blinowska. Granger causality and information flow in multivariate process. *Phys. Rev. E*, 70:050902, 2004.

[16] R. Brent and L. Lok. A fishing buddy for the hypothesis generator. *Science*, 308:504–506, 2005.

[17] S. L. Bressler and A.K. Seth. Wiener-granger causality: A well established methodolgy. *NeuroImage*, 58:323–329, 2011.

[18] D. R. Brillinger and A. Guha. Mutual information in the frequency domain. *Journal of Statistical Planning and Inference*, 137(3):1076–1084, 2006.

[19] S.G. Brush. History of the lenz-ising model. *Rev. Mod. Phys.*, 39(4):883–893, 1967.

[20] G. Burzaki. *Rhythms of The Brain*. Oxford University Press, Oxford, 2006.

[21] A. J. Butte and I. S. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5:415–426, 2000.

[22] D. A. Butts et al. Temporal precision in the neural code and time scales of natural vision. *Nature Letters*, 449, 2007.

[23] C. J. Cellucci, A. M. Albano, and P. E. Rapp. Statistical validations of mutual information calculations: Comparisons of alternative numerical algorithms. *Phys. Rev. E*, 71:066208, 2005.

[24] D. R. Chialvo. Critical brain networks. *Physica A: Statistical Mechanics and its Applications*, 340(4):756–765, September 2004.

[25] D. R. Chialvo. The brain, what is critical about it? *Amer. Institute of Phys.*, 28:1028, 2008.

[26] D. Chicharro and A. Ledberg. Framework to study dynamic dependencies in networks of interacting processes. *Phys. Rev. E*, 86:041901, 2012.

[27] K. Christensen and R. N. Moloney. *Complexity and Criticality*. Imperial College Press, London, 2005.

[28] B. A. Cipra. An introduction to the Ising model. *The American Mathematical Monthly*, 94(10):937–959, 1987.

[29] T. Cover and J. Thomas. *Elements of information theory*. Wiley, New York, 1999.

[30] G. A. Darbellay. An estimator of the mutual information based on criterion for independence. *Comput. Stat. Data. Anal.*, 32:1, 1999.

[31] J. Dufour and A. Taamouti. Short and long run causality measures: theory and inference. *Journal of Econometrics*, 154:42–58, 2010.

[32] T. E. Duncan. On the calculation of mutual information. *SIAM J Appl. Math.*, 19(1):215–220, 1970.

[33] T. E. Duncan. Mutual information for stochastic differential equation. *Information and Control*, 19:265–271, 1971.

[34] V. M. Eguiluz et al. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94(1):018102, 2005.

[35] I. Erb and A. Nihat. Multi-information in the thermodynamic limit. *Journal of Statistical Physics*, 115:949–976, 2004.

[36] P. Expert et al. Self-similar correlation function in brain resting-state functional magnetic resonance imaging. *J. R. Soc. Interface*, 2010.

[37] D. V. Foster and P. Grassberger. Lower bounds on mutual information. *Phys. Rev. E*, 83:010101(R), 2011.

[38] D. Fraiman, P. Belenzuela, J. Foss, and D. R. Chialvo. Ising-like dynamics in large-scale functional brain networks. *Phys. Rev. E*, 79:061922, 2009.

[39] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134, 1986.

[40] S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99:204101, 2007.

[41] K. Friston. Dynamic causal modeling and Granger causality comments on: The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. *NeuroImage*, 58:303–305, 2011.

[42] H.O. Georgii. *Gibbs measure and phase transition*. Walter de Gruyter & Co., Berlin, 1988.

[43] B. J. Gibb. *The Rough Guide to the Brain*. Penguin, London, 2007.

[44] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[45] C. W. J. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[46] C. W. J. Granger. Some recent developments in a concept of causality. *Journal of Econometrics*, 39:199–211, 1988.

[47] M. Harre and T. Bossomaier. Phase-transition-like behavior of information measures in financial markets. *Europhysics Letters*, 87:18009, 2009.

[48] T. Haruna and K. Nakajima. Permutation complexity via duality between values and orderings. *Physica D*, 240:1370–1377, 2011.

[49] T. Haruna and K. Nakajima. Symbolic transfer entropy rate is equal to transfer entropy rate for bivariate finite-alphabet stationary ergodic markov processes. *arXiv:1112.2493v2*, 2012.

[50] D. M. Hausman. The mathematical theory of causation. *Brit. J. Phil. Sci.*, 3:151–162, 1999.

[51] H. Herzel, A. O. Schmitt, and W. Ebeling. Finite size effects in sequence analysis. *Chaos, Solitons and Fractals*, 4:97–113, 1994.

[52] K. Hlavackova-Schindler et al. Causality detection based on information-theoretic approaches in time series analysis. *PhysicsReport*, 441:1–46, 2007.

[53] H. J. Jensen. *Self Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*. Cambridge University Press, Cambridge, 1998.

[54] H. J. Jensen. Probability and statistics in complex systems, introduction to. In *Encyclopedia of Complexity and Systems Science*, pages 7024–7025. 2009.

[55] J. Jeong et al. Mutual information analysis of the EEG in patients with Alzheimer's disease. *Clinical Neurophysiology*, 112:827–835, 2001.

[56] B. Julitta et al. Auto-Mutual Information function of the EEG as a measure of depth of Anesthesia. *33rd Annual International Conference of the IEEE EMBS*, 2011.

[57] A. Kaiser and T. Schreiber. Information transfer in continuous process. *Physica D*, 166:43–62, 2002.

[58] A. I. Khinchin. *The Mathematical Foundation of Information Theory*. Dover Publications, New York, 1957.

[59] J. L. F. King. Entropy in ergodic theory. In *Encyclopedia of Complexity and Systems Science*, pages 2883–2902. 2009.

[60] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[61] A. Kraskov, H. Stögbauer, and P. Grassberger. Hierarchical clustering using mutual information. *arXiv:q-bio/0311037v1*, 2008.

[62] W. Krauth. *Statistical Mechanics: Algorithms and Computations*. Oxford University Press, Oxford, 2006.

[63] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[64] W. Li. Mutual information functions versus correlation functions. *J. Stat. Phys.*, 60:823–837, 1990.

[65] Z. Li et al. Characterization of the causality between spike trains with permutation conditional mutual information. *Phys. Rev. E*, 84:021929, 2011.

[66] X. S. Liang and R. Kleeman. Information transfer between dynamical system components. *Phys. Rev. Lett.*, 95:244101, 2005.

[67] F. Lombardi et al. Balance between excitation and inhibition controls the temporal organization of neuronal avalanches. *Phys. Rev. Lett.*, 108:228703, 2012.

[68] M. Lungarella et al. Methods for quantifying the causal structure of bivariate time series. *J. Bifurcation Chaos*, 17:903–921, 2007.

[69] M. D. Madulara et al. Eeg transfer entropy tracks changes in information transfer on the onset of vision. *International Journal of Modern Physics: Conference Series*, 1(1):1–5, 2010.

[70] D. Marinazzo et al. Information flow in networks and the law of diminishing marginal returns: Evidence from modeling and human electroencephalographic recordings. *PLoS ONE*, 7(9), 2012.

[71] R. Marschinski and H. Kantz. Analysing the information flow between financial time series: An improved estimator for transfer entropy. *Eur. Phys. J. B*, 30:275–281, 2002.

[72] M. Martini et al. Inferring directional interactions from transient signals with symbolic transfer entropy. *Phys. Rev. E*, 83:011919, 2011.

[73] H. Matsuda. Mutual information of ising system. *Int. J. Theor. Phys.*, 35:4, 1996.

[74] R. Moddemeijer. On estimations of entropy and mutual information of continuous distributions. *Signal Processing*, 16:233–248, 1989.

[75] J. M. Nichols et al. Detecting nonlinearity in structural systems using the transfer entropy. *Phys. Rev. E*, 72:046217, 2005.

[76] A. Nihat. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–41, 2008.

[77] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 2008.

[78] M. Palus and A. Stefanovska. Direction of coupling from phases of interacting oscillators: An information-theoretic approach. *Phys. Rev. E*, 67:055201, 2003.

[79] A. Papana, D. Kugiumtzis, and P. G. Larsson. Reducing the bias of causality measures. *Phys. Rev. E*, 83:036207, 2011.

[80] J. Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[81] H. Peng. Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEE Transactions on Pattern Analysis and Machine Transactions*, 278:1226–1238, 2009.

[82] B. Pompe and J. Runge. Momentary information transfer as a coupling of measure of time series. *Phys. Rev. E*, 83:051122, 2011.

[83] K. J. Preacher and A. F. Hayes. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3):879–891, 2008.

[84] Q. Quiroga. *Quantitative analysis of EEG signals: Time-frequency method and chaos theory*. PhD thesis, Institute of Physiology and Institute of Signal Processing, Medical University of Lubeck, 1998.

[85] M. S. Roulston. Significant testing of information theoretic functionals. *Physica D*, 110:62–66, 1997.

[86] J. Runge et al. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Phys. Rev. Lett.*, 108:258701, 2012.

[87] N. Sauer. Causality and causation: What we learn from mathematical dynamic systems theory. *Transactions of the Royal Society of South Africa*, 65:65–68, 2010.

[88] T. Schreiber. Spatio-temporal structure in coupled map lattices: two-point correlations versus mutual information. *J. Phys. A*, 23, 1990.

[89] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[90] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142:346–382, 2000.

[91] C. E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[92] O. Sporns. Networks analysis, complexity and brain functions. *Complexity*, 8:56, 2002.

[93] C. Stamoulis, L. J. Gruber, and B. S. Chang. Network dynamics of epileptic brain at rest. *Conf. Proc. IEEE. Eng. Med. Biol. Soc.*, pages 150–153, 2010.

[94] M. Staniek and K. Lehnertz. Symbolic transfer entropy. *Phys. Rev. Lett.*, 100:158101, 2008.

[95] T. Suzuki et al. Mutual information approximation via maximum likelihood estimation of density ratio. *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory*, 1:463–467, 2009.

[96] T. Suzuki et al. Mutual information estimation reveals global association between stimuli and biological process. *BMC Bioinformatics*, 10:S52, 2009.

[97] E. Tagliazucchi, D. Fraiman, P. Belenzuela, and D. R. Chialvo. Criticality in large-scale brain FMRI dynamics unveiled by a novel point process analysis. *Front. Physiol.*, 3, 2012.

[98] D. Y. Takahashi, L. A. Baccalá, and K. Sameshima. Frequency domain connectivity: an information theoretic perspective. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pages 1726–1729, 2010.

[99] D. Y. Takahashi, L. A. Baccalá, and K. Sameshima. Information theoretic interpretation of frequency domain connectivity measures. *Biological Cybernetics*, 103(6):463–469, 2010.

[100] J. Theiler. Spurious dimension from correlation algorithms applied to limited time-series data. *Phys. Rev. A*, 34(3):2427–2432, 1986.

[101] M. Vejmelka and M. Palus. Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E*, 77:026214, 2008.

[102] R. Vicente et al. Transfer entropy: a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.*, 30:45–67, 2011.

[103] D. Welsh. *Codes and Crytography*. Oxford University Press, Oxford, 1988.

[104] G. Werner. Fractals in the nervous system: conceptual implications for theoretical neuroscience. *Front. Physiol.*, 1, 2010.

[105] R. T. Wicks, S. C. Chapman, and R. O. Dendy. Mutual information as a tool for identifying phase transition in dynamical complex systems with limited data. *Phys. Rev. E*, 75:051125, 2007.

[106] N. Wiener. *I am Mathematician: The later life of a prodigy*. MIT Press, Massachusetts, 1956.

[107] N. Wiener. The theory of prediction. In E. F. Beckenbach, editor, *Modern Mathematics for Engineers*. McGraw-Hill, New York, 1956.

[108] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 2008.

[109] R. M. Yulmetyev and P. Hänggi. Correlations in complex systems. In *Encyclopedia of Complexity and Systems Science*, pages 1615–1634. 2009.

Nomenclature

χ	Susceptibility
Γ	Covariance
Ω	$P($ condition fulfilled)
ρ	Correlation
D	Kullback Leibner
E	Expectation
H	Entropy
I	Mutual Information
L	Length of lattice
M	Magnetisation
N	Number of sites on the lattice
n_s	Number of states
P	Probability
$Q^{(\tau)}_{sgn(\gamma)}$	$P($ condition fulfilled given that $Z_{n-\tau} = \gamma)$
S	Sample size
s_X	State of site X

T Temperature

T_c Crossover Temperature

$T_{YX}^{(\tau)}$ Transfer Entropy of Y to X at time τ