University of Sussex

Information Transfer and Causality in the Sensorimotor Loop

James Thorniley

Submitted for the degree of Doctor of Philosophy in Informatics

University of Sussex

September 2015

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

James Thorniley

UNIVERSITY OF SUSSEX

JAMES THORNILEY, DOCTOR OF PHILOSOPHY

INFORMATION TRANSFER AND CAUSALITY IN THE SENSORIMOTOR LOOP

SUMMARY

This thesis investigates information-theoretic tools for detecting and describing causal influences in embodied agents. It presents an analysis of philosophical and statistical approaches to causation, and in particular focuses on causal Bayes nets and transfer entropy. It argues for a novel perspective that explicitly incorporates the epistemological role of information as a tool for inference. This approach clarifies and resolves some of the known problems associated with such methods.

Here it is argued, through a series of experiments, mathematical results and some philosophical accounts, that universally applicable measures of causal influence strength are unlikely to exist. Instead, the focus should be on the role that information-theoretic tools can play in inferential tests for causal relationships in embodied agents particularly, and dynamical systems in general. This thesis details how these two approaches differ.

Following directly from these arguments, the thesis proposes a concept of "hidden" information transfer to describe situations where causal influences passing through a chain of variables may be more easily detected at the end-points than at intermediate nodes. This is described using theoretical examples, and also appears in the information dynamics of computer-simulated and real robots developed herein. Practical examples include some minimal models of agent-environment systems, but also a novel complete system for generating locomotion gait patterns using a biologically-inspired decentralized architecture on a walking robotic hexapod.

Acknowledgements

This thesis would not have been possible without the valued assistance, insights and support of a variety of people. I would like to thank first of all my supervisors at Sussex, Phil Husbands and Andy Philippides.

Moreover, the members of the Centre for Computational Neuroscience and Robotics, as well as other staff and students at Sussex have helped immeasurably with advice, instructions, and indefinite patience with my less well thought-through ideas, not to mention the invaluable feedback of those who have made suggestions in response to the presentations of portions of this work given at Sussex: Lucas Wilkins, Thomas Greg Corcoran, Bill Bigge, Renan Moioli, Bruno Santos, Nick Tomko, Jose Fernandez, Matthew Egbert, Nathaniel Virgo, Simon McGregor, Lionel Barnett, Adam Barrett, and Keisuke Suzuki. Furthermore I would never have taken an interest in behaviour-based robotics if it wasn't for the suggestion of Jon Timmis.

Just as vital were the friends and colleagues who have provided moral support over the years. There are too many to mention everyone who has helped along the way – I thank you all. In particular my parents, Eric and Avis, and my brother and sister, Peter and Anne, for their patience. Special thanks to Sally Frampton for endless understanding, not to mention bringing me coffee just when I need it most.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Seldom do more than a few of nature's secrets give way at one time. It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information, entropy, redundancy,* do not solve all our problems.
>
> – Claude Shannon (1956)

In 1969 Clive Granger, building on a suggestion of Norbert Wiener, proposed a method for discovering causal influences based on temporal prediction (Granger, 1969): if a variable X can be used to predict future states of Y, above and beyond the prediction of Y that could be obtained from all prior information not including X, then X is said (in modern parlance) to *Granger cause* Y. Granger's application for this tool was in econometrics, but in recent years *Granger causality* and related approaches, notably *transfer entropy* (Schreiber, 2000), have proliferated in a variety of fields, including but not limited to the study of autonomous robotic systems.

A somewhat distinct but related development is the increasing application of information theory to the study of *complex systems*. A number of measures based on information theoretic concepts such as *entropy* (the level of "uncertainty" in a random variable) and *mutual information* (the extent to which two random variables are related to each other) have been used to describe the complexity of a physical process. Indeed there are a number of such proposals, going back at least to the *Kolmogorov-Chaitin* complexity, which is closely related to entropy for certain systems (Cover and Thomas, 2006), a notable more recent (and substantially different) model is the *statistical complexity* introduced by Crutchfield and Young (1989). Related information theoretic complexity measures have found particular application in characterising autonomous agents (Klyubin et al., 2005; Bialek et al., 2001; Der et al., 2008). Similarly, and perhaps in more direct

comparison to Shannon's oft-cited use of information theory to study digital and analogue communication systems (Shannon, 1948), such tools are often used to describe physical systems in *computational* terms – specifically to decribe features such as information storage and retrieval, communication and processing (Lizier, 2010).

The aforementioned transfer entropy measure is one of the primary objects of study for this thesis, and lies within the intersection of these two areas: as an information theoretic statistic, it is widely seen as a measure of complexity or computation, but it is also a generalisation of Granger's causality measure, and thus relates to causality in much the same way. However, both views of transfer entropy, as representing either causation or computation, have their own drawbacks.

Here we consider the application of transfer entropy to robotic locomotion (and to an extent autonomous robotic systems more generally). This offers a convenient and intuitive basis for explicating these approaches, and is used to develop a deeper understanding of them.

1.1 Robotics: Computation or dynamical systems?

The roots of our approach to autonomous robotics goes back at least to the work of Rodney Brooks (Brooks, 1986), and perhaps the "thought experiments" proposed by Valentino Braitenberg (Braitenberg, 1984) to describe how simple feedback connections could produce apparently autonomous behaviour. At its heart is a distinction between two radically different approaches to robot design.

What might be called "traditional" approaches to robotics regard the problem of robot control as requiring, in effect, the mathematical solution to a geometric puzzle (various introductions to these techniques are available, for example Craig, 2005; LaValle, 2011). A robot exists in a world defined by an n-dimensional state space, known as a configuration space or C-space. For example the position of each joint actuator, and the physical location of the robot (for a mobile system) constituting one dimension each. Many locations in this space are unreachable (certain combinations of joint actuator positions may be physically barred, for example in a robotic arm). In many cases movement through the space can only be achieved continuously or in certain directions. Further, the C-space maps onto the "workspace" – another space in which desirable outcomes for the robot can be defined. For example, a robotic arm must move its end-effector to a particular location: the location is specified in the workspace. The control system follows a methodology known as *sense-plan-act*:

- The system *senses* its current state in C-space. In some cases it may have to also "sense" its own problem definition (for example by sensing the position of an object to be manipulated

by a robotic arm).

- The system *plans* its solution to the problem: the problem specification (the workspace) must be mapped to the C-space. A valid path must be found between the robot's current position in C-space to its desired position.

- The robot must *act* out the plan – commands are sent to servo motors or other actuators such that the desired path through C-space is taken. After this, the problem has been solved.

There are multiple difficulties with this approach, identified by Brooks, particularly in the context of *autonomous* robots – where it is desired that the robot perform its tasks in unpredictable environments with minimal human intervention. In order to, for example, navigate to a particular location in a room, a *sense-plan-act* system must either be programmed with, or create from sensor data (and complex computational analysis) some kind of internal *map* or computational representation of the room. If someone moves a table in the room, a robot with a pre-programmed C-space will attempt to pass through the table (thinking it is a valid trajectory when it no longer is), or must expensively recalculate its internal map and re-plan its solution.

These problems led autonomous robotics research to investigate radically different approaches to solving problems in unpredictable environments. *Behaviour-based* robotics (Arkin, 1998) – a term used to describe robotic systems designed in the vein of Brooks (1986) – emphasises the use of combinations of very simply defined behaviours. For example, it is in fact quite straightforward for a mobile robot to avoid obstacles immediately in front of it, by simply turning away when forward proximity sensors report a close by object. This simple behaviour can be "mixed" with other behaviours – navigating towards a light for example, to produce an effective solution with relatively low computational costs. *Evolutionary* robotics (Floreano et al., 2008; Harvey et al., 2005) takes a slightly different tack – again the sense-plan-act cycle is discarded, this time in favour of developing a controller by specifying a class of possible controllers, then generating a population of candidate controllers from within that class which are then evaluated by a specified *goal function* (measuring the success of the robot in achieving its task). Those controllers which achieve less optimal results according to the goal function are discarded, and those which are more successful are maintained, duplicated, randomly mutated and sometimes combined in a process somewhat analogous to biological evolution. This method can be used to find controllers which solve complex problems in ways not obvious to human designers.

What these approaches have in common, in contrast to the sense-plan-act methodology, is the absence of internal representations or maps of the problem space from which computational algorithms calculate solutions. Rather, they emphasise that the robot interacts continuously with

its physical environment, often purely "reactively" (without even any explicit internal memory). In Brooks's words, "the world is its own best model" (Brooks, 1990). This attitude has been described as the "dynamical systems" approach to robotics (Beer, 1995; Harvey et al., 2005) – autonomous robots are viewed primarily as physical systems which respond in defined ways to perturbations or external stimuli, rather than as computational systems which calculate and plan solutions to problems defined using internal representations. It is worth pointing out that there is a related body of work which aims to characterise human and animal cognitive function in similarly dynamical systems, or non-computational, terms (Clark and Toribio, 1994; Van Gelder, 1995; Chemero, 2009).

1.2 Synchrony in robotic gaits

Following this approach, we will treat a robotic gait generation system in dynamical systems terms. A consequence of this is that we will have particular interest in the phenomenon of synchronisation. This is a well-studied feature of physical systems (for a general discussion see Pikovsky et al., 2001), and has been used to model flocking and schooling in birds and fish, biological rhythms such as circadian rhythms and the synchronous behaviour of fireflies to name a few examples (Strogatz, 2004). Furthermore, synchronisation has been hypothesised to play a role in animal and robot gait generation – gaits incorporate patterns of spatio-temporal symmetries which can be generated by symmetrically connected networks of coupled oscillators (Collins and Stewart, 1993). This paradigm has been used to develop models of *central pattern generators* (Ijspeert, 2008) – neural systems which use synchronisation to coordinate motor patterns of limbs or other actuators. For this reason, synchronisation (in the general sense of coupled dynamical systems which mutually influence each other to produce similar dynamics) will be a substantial area of study for later chapters in this thesis (including those which do not directly address gait generation).

Dynamical synchrony typically arises when one oscillatory system exerts a small but significant influence on another (the influence may be one-way or mutual depending on the circumstances). As a result, small deviations between the two systems produce forces which act to bring the two trajectories in line with each other (see figure 1.1).

Synchrony presents a particular problem for the use of transfer entropy. Mutually synchronised oscillators may trace out identical trajectories, meaning that although they are interacting, the past states of the two systems are identical. This means that the state of a driving oscillator cannot improve the prediction of a driven oscillator beyond the prediction obtained from the past of the driven oscillator alone, since the two states are identical. In other words, the information from the cause is *non-separable* from the information about the past of the effect. This is one of the main

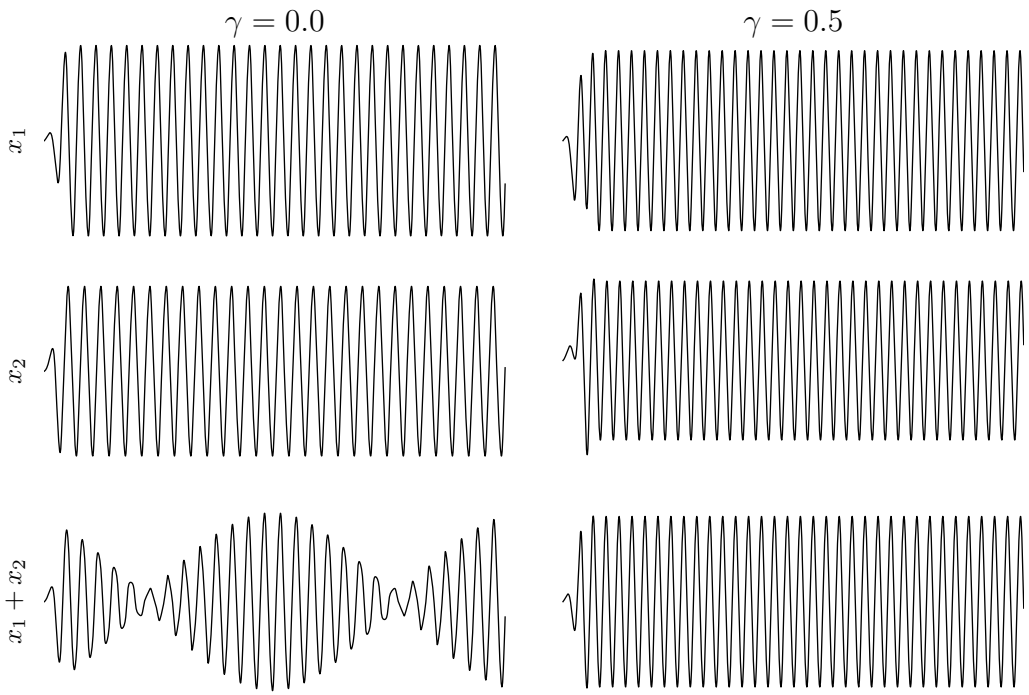


Figure 1.1: Dynamical synchronisation. An example of this phenomenon involving two simple oscillators can be seen in the system given by the differential equations

$$\begin{aligned} \ddot{x}_1 &= q_1\dot{x}_1 - x_1 - 0.1\dot{x}_1^3 - \gamma x_2 \\ \ddot{x}_2 &= q_2\dot{x}_2 - x_2 - 0.1\dot{x}_2^3 - \gamma x_1 \end{aligned}$$

Where q_1, q_2 and γ are fixed parameters and dots represent differentiation with respect to time t. The first three terms of each equation define a self-sustaining oscillation provided q_1 and q_2 are positive. That is, with γ set to 0, x_1 and x_2 are the states of distinct (uncoupled) dynamical systems. If we set q_1 and q_2 to different values (in this example we use $q_1 = 1$ and $q_2 = 1.5$), the two systems will oscillate at slightly different frequencies. This is illustrated by the time series traces in the left hand column, generated by numerically integrating these equations with the starting value of $x_1 = x_2 = 0.1$ – though x_1 and x_2 are similar, the latter oscillates at a slightly lower frequency. The different can be seen more clearly in the trace of the sum $x_1 + x_2$, which "beats" as the two oscillators run in-phase and out of phase with each other over time. However, if we set $\gamma = 0.5$ (right hand column), the value of x_2 influences the changes in x_1 and *vice versa* – as a result, the two oscillators stay in-phase with each other producing a constant amplitude trace when we add the two signals together.

difficulties with the interpretation of transfer entropy as a measure of causation.

1.3 Information dynamics in agents

The dynamical systems approach (more than, say, purely stochastic models that are more closely comparable to communication systems) also presents a problem for the interpretation of transfer entropy as a measure of computation – as discussed, we are interested in robots which do not have explicit internal representations of their problem space, and do not calculate solutions from such a representation. Thus in a classical sense, such systems are often argued to be non-computational (Chemero, 2009). On the other hand, this boundary is blurred somewhat by the *morphological computation* approach advocated by Pfeifer et al. (2006). This refers to the capacity of the body itself to perform a kind of computation, for example one can grip an object by simply contracting one's fingers around it – without explicitly calculating the correct configuration of each joint. This makes use of the inherent compliance of human joints and muscles to "offload" computation that might otherwise have to be performed in the brain to the body's own morphology.

This concept of morphological computation is more reminiscent of Brooks's view that the "world is its own best model" than the classical sense-plan-act paradigm. In fact, the concept of information itself has proven useful in describing this type of computational "offloading". Informally, we can think of Brooks's autonomous robots as acquiring information about the world through constant sensor feedback, almost as a *substitute* for an accurate internal map – a correct internal map is harder to come by practically speaking, but the point is that the robot must *have access to* certain information to achieve its goals, whatever way such information is obtained. More formally, Touchette and Lloyd (1999) showed that the maximal amount by which a control system can reduce the entropy of the system it controls is limited by the information it receives about that system. This argument is very similar to that made in an earlier paper entitled "Every good regulator of a system must be a model of that system" by Conant and Ashby (1970). A "regulator" here refers to a system which keeps another system in a certain desired or "good" state by perturbing it in just the right way to combine with external perturbations to the target system. The point made by Conant and Ashby is that as the complexity of the system being regulated increases, the complexity of the regulator must also increase to match this. If there are more ways that the system can respond to external perturbations, then there must be more ways that the regulator can react to and counteract those responses. This argument formalises a relationship between information and control that has inspired the development of autonomous systems from information theoretic principles (Klyubin et al., 2005; Zahedi et al., 2010; Der et al., 2008). There has also been a great deal of interest in "quantifying" or otherwise studying what is often termed *information dynamics* – i.e.

investigating the connection between information relationships and the behaviour of autonomous systems (Lizier, 2010; Lungarella and Sporns, 2006; Moioli et al., 2012; Schmidt et al., 2012; Zahedi and Ay, 2013).

1.4 Inference of causation

Granger's proposal of a measure of temporal causation has long been seen as "not really causation" – hence the introduction of the language "X Granger causes Y", rather than simply "X causes Y". We have already noted one reason for concern, namely the issue of *non-separability* (at the end of section 1.2).

Despite this and other practical problems, not to mention a host of thorny metaphysical issues, causation has recently become a major field of study. A notable development has been the graphical modelling approaches based on directed acyclic graphs (DAGs) known as *causal Bayes nets* (Pearl, 2009; Spirtes et al., 2001). These enhance a concept of causation based on probabilistic association, derived from Reichenbach's *principle of the common cause* (Reichenbach, 1956) – loosely stated, a statistical association between to variables must arise either because one causes the other or there is a third factor which causes both. However they have also allowed for introducing causal concepts beyond this – Pearl in particular introduces the *do()* operator, which represents "intervention" in a causal system. This bears close resemblance to the "manipulability" account of causation (causes are those things which can be altered in order to change their effects, Woodward, 2004).

Granger's approach is also closely related to the principle of the common cause (Holland, 1986), and by extension so is its information-theoretic equivalent transfer entropy. The literature reveals a number of objections to the various accounts of causation in terms of the principle of the common cause, causal graphical models and the Granger method (Sober, 1984, 2001; Cartwright, 1994; Dawid, 2009). Nonetheless, Granger causality and transfer entropy have recently been applied in a number of fields including neuroscience (Vicente et al., 2011; Seth et al., 2011), economics (Marschinski and Kantz, 2002), climate physics (Runge et al., 2012b) and artificial life (Schmidt et al., 2012).

Graphical causal modelling in particular has inspired many recent developments in causal inference, including variations on transfer entropy (Runge et al., 2012b) and indeed alternatives to transfer entropy aimed at solving its apparent drawbacks such as non-separability (Ay and Polani, 2008; Janzing et al., 2013; Sugihara et al., 2012).

1.5 Thesis overview

This thesis is about causality in the sensorimotor loop – it addresses the questions of what we mean by causality and what its relevance is for agent behaviour. It aims primarily to explicate and study the apparent problems with treating transfer entropy and related information theoretic statistics as "measuring" or "quantifying" causation. This is achieved through a combination of theoretical analysis (part I of the thesis) and practical examples on autonomous simulated and real robotic systems (part II).

Beginning with part I, chapter 2 addresses the topic of causation in detail. It gives an overview of recent developments and debates in this area, aiming to translate them into terms of particular relevance when we are looking at autonomous agents and complex systems. There are a number of alternative approaches to causation which have been suggested by a variety of authors, and this chapter describes various relationships and important distinctions between them. It focusses in particular on the probabilistic and Bayes nets approaches that form the basis of much of the theory used later in this thesis. It does not, however, discuss in great detail transfer entropy or Granger causality – these are introduced subsequently and seen as variants of the probabilistic causation introduced in chapter 2, rather than a separate entity.

Chapter 3 serves several functions. Within the context of the thesis it is the first example of an empirical use of transfer entropy that will provide some context for later applications. However, its novel contribution is the proposal of a distinct *method* for calculating transfer entropy based on examining the phase data of oscillatory systems alone, rather than by incorporating the entire state of the oscillator (i.e. it discards amplitude information). It is argued that where we are interested in synchronising systems, looking at the phase information preserves the aspects we are interested in and discards the irrelevant amplitude dimension, thus reducing the complexity of the data used in the transfer entropy calculation.

Chapter 4 sharpens the formal approach to causality taken here. It constructs from first principles the relationship between singular causal hypotheses (e.g. X causes Y) and information theoretic statistics in the context of a given background network of "known" causal relationships modelling using Bayes net methods. This presents an account of the relationship between mutual information and causal networks that explicitly models the *inferential* role of the information statistic. We argue that it is crucial to maintain a distinction between measurement of causal *strength* and *complexity*. By analogy with the Neyman-Pearson model of statistical testing, we see that information can be used to evaluate something akin to the epistemological clarity of a particular causal hypothesis. The model introduced in this chapter is further extended to the context of time series and it is shown how transfer entropy fits within the more general model of causal inference.

A significant aim of the first part of this thesis is to critically evaluate claims of potentially superior (and at least distinct) approaches to causal inference other than transfer entropy. Chapter 4 addresses one proposal, the information *flow* proposed by Ay and Polani (2008). Furthering this project, chapter 5 considers the convergent cross-mapping proposal of Sugihara et al. (2012). This alternative is based on somewhat different foundations than transfer entropy, however chapter 5 demonstrates that it is closely related to a form of mutual information. In fact, as we show, it can therefore be subsumed in the same information theoretic framework used to describe transfer entropy. This provides a beneficial and novel analysis of the relationship between information theoretic and cross-mapping based approaches.

The closing chapter of part I, chapter 6 returns to the subject of the distinction between *strength* and *inference* based understandings of information transfer. We consider the relationship between the new outlook proposed in chapter 4 to *information dynamics*, particularly in the context of autonomous agents. This chapter introduces a distinct phenomenon of such information dynamics that we can predict in light of the earlier discussion – namely that of "hidden" information transfer. Certain patterns of synchrony can enable information to pass from one node to another (physically) very far away, with the information not manifesting in intermediate nodes, even though it must pass through them. In other words, we can have a chain of causal influences, where A affects B which effects C, but where the causal influence is manifested by high information transfer (is "epistemologically clear") between A and C, but not between A and B. We demonstrate how this phenomenon appears in a theoretical model of communication through synchronised chaotic systems, and predict that it may also occur in agent-like physical systems.

In part II, we aim to apply the theory and concepts developed in part I to robot control, particular gait generation for a legged robot. We start by reviewing some important topics in the design of gait generation systems in chapter 7. There are a number of possible techniques, and this chapter describes in brief the comparisons and reasons for taking the approach we do. We investigate approaches that incorporate dynamical synchronisation as a tool for generating spatio-temporal patterns of symmetry, and particularly wherein feedback from the environment and body plays a role in determining the overall behaviour.

We start with a minimal example in chapter 8. This chapter is the first identified example of the hidden information transfer phenomenon in an embodied robotic system. In this minimal example, the true causal roles of the various elements of the system can be made explicitly clear with relatively little difficulty. This contributes a significant demonstration of the relevance of the hidden information transfer concept in an embodied system, and indeed provides a clear example in part by virtue of the relative simplicity of the system under study.

More complex examples are developed in chapter 9 and 10. Here we develop a distributed architecture for gait generation in a real hexapod robot in parallel with a closely matched computer simulation. Our gait generation architecture, though based on similar models such as Walknet (Cruse et al., 2002), represents a novel engineering approach making use of position sensing in the robot's joint servos to provide physical feedback to the distributed controller systems. As a result, the architecture presented has the unique ability to generate a stable tripod gait without any internal (electronically mediated) coordination between the left and right-hand limbs – instead, synchronisation can be achieved purely through mechanical coupling along the contralateral axis.

The information theoretic analysis of causation in this system is presented in chapter 10. We develop here a set of statistical tools based on our understanding of information as a tool for inference, giving a more principled way to interpret transfer entropy results taken from a single experiment. This allows us to again investigate the presence of hidden information transfer, this time in the context of a more sophisticated and multi-faceted network of causal influences. Here we have two primary channels for causal influences to pass between the several limbs of the robot: mechanical (through the body and ground) and electrical (internal connections in the control system). We investigate how changes in both these coupling systems affect the recorded transfer entropy between variables affecting limb oscillators.

1.6 Original contributions

The major contributions of this thesis are summarised here:

- We introduce two novel tools based on theories of causal influence: first the stroboscopic transfer entropy (chapter 3) is a variant of transfer entropy based purely on the phase relationship data in a time series. Second the cross embedded mutual information (chapter 5) is an information-theoretic tool based on the same theory that underlies the (regression-based) convergent cross-mapping approach.

- The development of a well-founded and novel theory of the use of information transfer as a tool for inferring the presence of causal influences. While the idea that information may be an indicator of causation has been considered by many authors, previous proposals have tended to either adopted a naive interpretation of an inferential statistic as akin to something that measures *strength* of causation – an approach shown here to have intrinsic failings – or to simply hedge causal claims and side-step any formal justification. What is needed is the principled connection of statistical inference techniques and the metaphysical theories underlying causal Bayes nets introduced in chapter 4.

- We describe a general framework of causal inference and show how existing proposals for measuring causal influences (information flow, chapter 4, and convergent cross-mapping, chapter 5) can be situated within this theoretical framework, alongside our formal causal interpretation of information transfer. This significantly adds to our understanding of all approaches by allowing direct comparison according to a theory which applies equally to all methods.

- This thesis identifies the concept of "hidden information transfer" as a very general schema of information dynamics, and proposes certain scenarios where we might expect this phenomenon to occur in autonomous agents. Chapter 6 introduces the theoretical basis of this, practical demonstrations are given in chapters 8 and 10.

- We introduce a novel hexapod gait generation algorithm as a test-bed for the statistical techniques used in this paper. This algorithm is capable of producing multiple gaits through emergent synchronisation between distributed oscillators, and demonstrably incorporates information feedback through mechanical coupling. This is introduced in chapter 9.

According to the interpretation I will advance here, statistics such as transfer entropy are well understood as tools for *causal inference*. This means that rather than seeing higher transfer entropy as indicating higher causal strength, we should see it as a way of testing particular causal claims. Thus the overarching theme of this thesis is the juxtaposition of two views of information and causation. The first we can call the *causal strength* (CS) interpretation, wherein we look for information theoretic tools which will measure strength of causation. The second is the *causal inference* (CI) approach that I am advocating here. The goal of this thesis is to provide a convincing argument that the latter is a substantially different and useful alternative to the former. Not only is it more strongly philosophically justified, but we will see that it has a number of specific advantages over the causal strength interpretation from the point of view of understanding causation in the sensorimotor loop.

1.7 Glossary of terms

A number of terms are used in the literature in both specific technical as well as "informal" senses. In this thesis it should be clear where a term has a technical meaning and where it is used in the more informal sense. This section will help to clarify some particularly important points from the outset.

Information	Informally, the precision or certainty about a variable or state provided by another variable or state. Alternatively, might refer to a richer concept of meaning or utility attached to some data. Formally, this thesis is only concerned with definitions based on probabilities, specifically *mutual information* as defined by Shannon (1948), or the related *conditional mutual information*. The "information" defined by Kullback (1959) is referred to here in common with most modern literature as *Kullback-Leibler divergence*.
Information transfer	Informally, any probabilistic description of the influence (causal or merely correlative) of the state of one variable on the state of another, typically over time. Formally, defined in chapter 4 as a conditional mutual information satisfying particular conditions with respect to a causal graph making it suitable for the evaluation of a causal hypothesis.
Information flow	Informally synonymous with information transfer (we avoid this use). Formally, defined by Ay and Polani (2008) again with respect to causal graphs but also considering interventions.
Complexity	An informal term describing systems whose behaviour is in some sense difficult to predict in spite of not being (entirely) random, such systems may well be defined in terms of relatively simple structural components.

Stability — Informally often refers to systems with have static or thermodynamic-equilibrium dynamics (i.e. the state of the system, or the distributions of states of an ensemble, is unchanging over time) – this use is avoided here. There are two technical uses here: in robotics, refers to a gait or other behaviour which can persist over time without catastrophic failure (e.g. a robot would be unstable if it fell irrecoverably onto its side whilst walking). In dynamical systems theory, *Lyapunov stability* refers, technically, to volumes in a dynamical system which shrink over time (Strogatz, 2000). An intuitive way to think of this is that a *stable attractor* is a region of state space within which all trajectories that start in this region will remain in the same region, generally approaching a smaller and smaller subset of the region. Note in particular that for example repetitive cyclic and chaotic dynamics may be stable by this definition but not by the more informal definition described above.

Ergodicity — Has multiple technical definitions depending on context. Notably, an ergodic Markov chain is a Markov chain which has no periodic cycles or disconnected subgraphs. Ergodicity for dynamical systems is defined with respect to *ensembles* of dynamical systems (a set of solutions to a dynamical system over which a probability measure is given), and specifies conditions under which the statistics calculated by averaging over time series data taken from one member of the ensemble converge to the equivalent statistic taken over the entire ensemble at any particular point in time. Formal definitions are given in Shannon (1948), see also further discussion in chapter 4. Ergodicity should not be confused with *stationarity* which refers to systems which behave similarly over time. Stationarity is a necessary but not sufficient condition for ergodicity.

Part I

Theory

Chapter 2

Causation and inference

How do we account for the behaviour of an artificial or biological agent? A common view is that agents can be seen as *systems*, that take an input (in the form of sensor data), perform some processing and produce an output (motor actions). A recent trend has been to characterise such a systems approach in specifically *causal* terms – changes in inputs *cause* changes in outputs, via networks of causal connections within the "processing" component. It is also common to find talk of agents obtaining and using *information* – which may mean information about the "outside world" obtained via the sensors, or perhaps the information shared between social animals for example. This thesis focuses on the connection between causation and information, and so to begin with, this chapter will introduce the necessary background to develop a coherent "causal" view of agent behaviour.

Most of the literature surveyed in this chapter addresses causation from the perspective of either philosophy of science – applied often to economics, social sciences and epidemiology – or machine learning and artificial intelligence. The literature generally does not pertain directly to robotics or agent behaviour (though I will aim to discuss concepts using simple robotic systems as examples). This is simply where the literature on causation has historically been targeted – the application of these ideas to agent behaviour seems to be a relatively new phenomenon. The reason for introducing this background is to establish a firm description of causation that is consistent and will be applicable in later chapters. Without this foundation, one is liable to get confused by differing intuitions and interpretations around these somewhat complex (and sometimes controversial) topics. Much of the difficulty with interpreting the extant literature that *does* directly discuss agent behaviour (which we will come to in the next chapter) arises from the fact that in that area, the concept of causation is often loosely and quite poorly defined.

As we will see, causation can usefully be thought of in a number of different ways. This chapter will outline the approach that I will adopt and develop in this thesis, which consists of a

dual system of *physical models* and *Bayes nets*. This chapter shows how this system relates to the common approaches to causation found in the literature.

2.1 Why complex systems confound "naive" causation

To some, it may seem surprising to insist on elaborating substantially on the question of causation. Although it is uncontroversial to distinguish causation from "mere" correlation, many argue that modern scientific practise has more-or-less resolved the question of how to find out about causes. Specifically, X is a cause of Y just when changing X, while keeping everything else the same, results in a change in Y (Woodward, 2004, ch. 2). All we need are methods that establish causal connections by this logic.

The prototypical modern method for determining such relationships is the *randomized controlled trial* (RCT). This is an experiment where a number of subjects (e.g. human participants in a psychology experiment or drug trial) are divided by random assignment into two or more groups. Different groups are given different treatments, and an outcome measure is taken for each subject. One can then determine statistically whether the outcome measure varies between each group (for example by comparing group means, though the statistical techniques are usually more involved). The technique aims to satisfy the condition of keeping "everything else" (other than the treatment) identical between the groups, the randomization being a key element of this. Thus if the outcome changes between groups, we can infer that the treatment is a cause of that outcome.

This method is in some senses a relatively recent invention. Its earliest roots are often traced back to James Lind (1716-1794), sometimes called the "father of the clinical trial" (Twyman, 2004). Lind experimented on various different methods of preventing scurvy among sailors, but he introduced only one significant aspect of the RCT – namely the introduction of various control groups for comparison. C. S. Pierce (1839-1914) is credited with introducing *randomization* in a series of psychological experiments conducted in the 1880s (Stigler, 1978), though this aspect of his work seems to have been largely forgotten for some 30 years until R. A. Fisher (1890-1962) developed the modern form of the RCT (Hacking, 1988). Fisher's 1935 book *The Design of Experiments* (Fisher, 1935) introduced the RCT to a wider audience, and from there it has come to be seen as "the most rigorous scientific method for evaluating the effectiveness of health care interventions" (Akobeng, 2005), the "gold standard" (Sackett et al., 1996; Akobeng, 2005; Henderson, 2012, p. 129) and the "best way of determining if a policy is working" (Haynes et al., 2012).

Thus it is tempting to stop at that and take the ideal, well-conducted RCT as *all there is* to causation. Colquhoun (2011) describes the Fisherian RCT as "the essential underlying condition

of causal inference... Everything else is worse." From such a perspective an extended discussion of the philosophy underlying causation seems rather pointless – certainly, there are cases where RCTs are impractical (because we physically cannot intervene on the relevant variables) or simply have not been done *yet*, and thus we might have to rely on the "worse" inferences we might make from observational studies. Nonetheless the (well-conducted) RCT is still the ideal goal of causal inference.

There is plenty of criticism to be found of what one might call "RCT exceptionalism" in the vein of Colquhoun. But this typically focuses on questions of *external validity* (the applicability of results obtained in highly-controlled experiments in untested situations) and the sometimes overlooked importance of observational evidence (Cartwright, 2007a; Cartwright and Munro, 2010) – while that may be important in many areas such as social science and economics, it is not a key problem in the current context. For our purposes, the inability to intervene is hardly a problem – this thesis is primarily focussed on artificial systems, where we have essentially a limitless capacity to modify and tweak the systems under study, simply by altering variables or code in a computer program.

Rather, the problem for us with attempting a naive interpretation of causation is the tendency of *complex* systems to confound the straightforward approaches to causation. By "complex systems", I mean systems characterised by networks of causal influences including feedback loops that dynamically evolve in non-linear ways.[1] Section 2.3 discusses various cases, including a relatively basic robotic example, that do not fit the simple model of cause and effect. The interventions we make to causal variables do not have to *always* change their effects, and in a complex system we cannot *a priori* know how any given causal relationship may be contingent on background factors being arranged in just the right way.

The chapter will build towards a formalisation of causal systems based on the concepts of physical models and causal Bayes nets. I aim to distinguish what we can rely upon in these theories, and what we must be more cautious about, especially in the context of complex systems.

2.2 Control, regularities and counterfactuals

One of the primary reasons for being interested in causation is control of some sort or another: "causal and explanatory claims are informed by our interest as practical agents in changing the

[1] There is no precise definition of a "complex system" (Ladyman et al., 2013). I use the term in this chapter in a loose sense, but primarily to mean *dynamical* rather than *compositional* complexity as described by Kuhlmann (2011) – systems characterised by emergent dynamical features that arise from non-linear interactions, rather than those that are complex by virtue of having many "fine details" to their description.

world." (Woodward, 2004, p. 25). In the context of robotics, the importance of control is clear: perhaps we want to make a mobile robot change direction, which part of the system should we modify to achieve this? What external stimuli would cause particular changes in behaviour? Presumably that part of the system which we can modify to effectively control the robot's heading could reasonably be called a cause of the overall heading.

On the other hand, causal explanation could be viewed simply as an important component of scientific understanding, it is not necessarily the case that one is practically able to use it for any sort of control. Perhaps *in principle* some control or intervention could be made, but we might want to say for example what area of the brain is 'responsible' for some higher level behaviour without ever intending to actually change that behaviour by (directly) influencing that brain region. These two types of cause are nonetheless similar, in that they focus on general properties: causal influences which are "always" present.[2]

Dawid (2000) calls general causal descriptions of this sort *Effects of Causes* (EoC), as opposed to specific questions about what event caused what other event: *Causes of Effects* (CoE). Dawid gives by way of example the distinction between the questions, "does aspirin cure headaches?" (a general EoC) and "did the headache I had go away as a consequence of taking an aspirin?" (CoE). Answering these different questions requires very different thinking though they are sometimes mixed up: indeed there are ways in which we can form at least loose connections between the two.

Cause-as-general-regularity (or EoC in Dawid's taxonomy) is a concept associated with the Scottish philosopher David Hume (1711-1776). Hume was an empiricist who argued that causal relations cannot be directly observed – all we have access to is the observation of a recurring association. Observing a (putative) cause in conjunction with its effect (many times over) is what leads us to believe that it *is* the cause.[3]

Specific questions or CoE are often viewed in terms of *counterfactuals* – for example A is the cause of B when "if A had not occurred, B would not have occurred". This type of approach is usually traced back to John Stuart Mill (1806-1873) (Menzies, 2009):

> Thus, if a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the cause of his death. (Mill, 1843, ch. V, §3)

[2] "Always" is in quotes because as has already been discussed, most causal relationships are contingent. Usually "always" in this context would be understood to mean "whenever relevant background factors are organised appropriately".

[3] Hume covers the subject extensively in *An Enquiry Concerning Human Understanding* (Hume, 2006, original posthumous publication 1777), especially sections IV to VII. See also Simon (1977); Morris (2013); Winship and Sobel (2004).

Later, counterfactual logic was adopted by statisticians as a way of analysing causal effects. Jerzy Neyman[4] introduced a mathematical model of counterfactual analysis, which became well known after being developed by Donald Rubin and Paul Holland in the 1970s and '80s (Rubin, 1974; Holland, 1986; Winship and Sobel, 2004). It is now usually known as the *Neyman-Rubin* model.

The model can be introduced as follows. Imagine we have a population of simple phototactic robots – two wheeled agents which navigate towards the light (the exact mechanism is unimportant for now). Call the population $\mathcal{U}$. Each robot in the population, $u \in \mathcal{U}$, is experimented on separately – at the start time of a given experiment, call it t_1, the current robot u is set to a pre-defined starting state – initial position and heading – identical for all experiments. We then apply one of two possible stimuli, either L – a light on the left hand side of the robot or R – a light on the right hand side. The two possible stimuli produce two potentially different outcome functions Y_L and Y_R which map from an instance u to some outcome measure taken at a later time t_2. For example, the heading of the robot relative to its starting heading a second after the initial stimulus. Thus if we measure angles as increasing for clockwise rotation, perhaps we have:

$$Y_R(u) = 10^\circ$$

$$Y_L(u) = -10^\circ$$

That is, if we present a light on the front-right the robot turns 10° clockwise in a second, and correspondingly for the left hand stimulus it will turn 10° anti-clockwise. If this model is accurate then it would appear that we do indeed have a successful phototactic robot. The *causal effect* of changing from the left-hand to right-hand stimulus is defined simply as:

$$Y_R(u) - Y_L(u) = 20^\circ$$

However, we run up against what Holland (1986) calls the *fundamental problem of causal inference*. This is that we cannot observe $Y_R(u)$ and $Y_L(u)$ at the same time for any *particular* member u of the population $\mathcal{U}$. Of course in the case of these simple robots there seem to be obvious ways to get around this: the absolute time (we suppose) is probably not important, so we could just present R for one second to measure Y_R, then take it away and present L to find Y_L. But in principle the battery might run out as soon as the first part of the experiment is finished, so in the second part we would not see any movement and would measure a 0° change in heading. In any case, the robot may have noisy sensors and imprecisely controlled motors, so maybe it doesn't turn exactly 10° per second: sometimes it turns 10°, sometimes 11°, sometimes 9°. So if we use

[4]Neyman (1923) (original in Polish, a translation appears in Dabrowska and Speed, 1990)

the one-after-the-other strategy, and present the left hand light at $t_3 > t_2$, the robot still might move slightly differently to how it *would have moved* if we had presented the left hand light at t_1.

The point is that we can never *observe* the response of a given unit to both of the stimuli – since in reality we must choose to present one or the other. We can however use some sensible assumptions and a bit of logic to make certain inferences. If, for example, the robots all behave in approximately the same way, then surely we could just assign one subset of the population $\mathcal{U}_R$ to receive stimulus R and the other $\mathcal{U}_L = \mathcal{U} \setminus \mathcal{U}_R$ to receive L. We have already said that there is some random perturbation specific to each trial, but across this large population that can be averaged out. The differences of the mean outcomes

$$\frac{1}{N_R} \sum_{u \in \mathcal{U}_R} Y_R(u) - \frac{1}{N_L} \sum_{u \in \mathcal{U}_L} Y_L(u)$$

can be seen under some reasonable assumptions to approximate the central tendency or *average causal effect*, defined as the expectation of the difference between the treatment outcomes across the population.

$$E_{\mathcal{U}}[Y_R(u) - Y_L(u)]$$

This takes us quite easily from the Neyman-Rubin model back to a general regularity or EoC: the average causal effect quantifies something that (essentially by assumption) is broadly the same across the population. If both stimuli cause the robot to turn in the appropriate direction, on average, by 10° (but plus or minus in each case some random perturbation with zero mean), then the average causal effect (comparing L to R in totality) is clearly 20° and we should observe this from a large trial of similar robots.

Yet for causes of effects – specific questions about particular robots – we still have not really answered the question of what *this* robot would do if it had been presented with something other than what it was in fact shown. If we let one robot u receive stimulus R and see it turn 11° to the right, how do we know what it would have done if we presented L instead? Perhaps the 11° turn was the sum of a 10° deterministic effect and a 1° influence from "random noise" (the latter could in principle have been 0° or -1° by chance, but in the current instance it was 1°). Thus, perhaps we think that if the stimulus L had been presented instead only the deterministic effect would have been changed (to -10°) and the random component would have remained "the same", to give a net heading of -9°. In other words, we take $Y_R(u)$ and subtract the average causal effect to determine $Y_L(u)$.

However, one could surely argue that by its nature, the actual random perturbation has not been determined for the case of the untested stimulus. It is not obviously any less reasonable to

argue that the response $Y_L(u)$ (which has not been observed) could still in principle take one of several values depending on how the random perturbation would have come about. Although we can find the average causal effect by testing different robots with either L or R, we can never answer this question about whether, if we swapped the treatment given to a particular robot, the random influence would be exactly the same. This objection is one Dawid (2000) raises to the use of counterfactual models such as this. He points out that two statisticians could disagree on whether the noise added under the L condition is the same as that added under R[5] and have no empirical way to resolve the question.

Dawid views this question of the relation between the noise under the two treatments as "metaphysical" and inherently undecidable. On the other hand, Pearl (2000) compares Dawid's objections to what he calls "red scares" about metaphysics in statistics. Karl Pearson (1857-1936) is quoted by Pearl: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amid the inscrutable arcana of modern science, namely, the category of cause and effect" (Pearson, 1911). Pearl agrees that there is a real distinction between EoC and CoE questions ("does aspirin cure headaches" versus "did the aspirin I took cure my headache"), but challenges Dawid "to express [CoE], let alone formulate conditions for its estimation in a counterfactual-free language" (Pearl, 2000, p. 429). Pearl also claims that the conditions needed to answer CoE questions can not only be formulated but empirically tested.

To summarise: EoC questions can potentially be formulated in a counterfactual language (the above definition of *average causal effect* started with counterfactual notation but resulted in describing a general regularity), but some authors such as Dawid argue counterfactuals are to be avoided. CoE questions, it seems, are very difficult to formulate without recourse to some kind of counterfactual language: for Dawid this is a major problem, for Pearl it is not. For our purposes what can be agreed is that EoC questions are less problematic, since whether we describe them with counterfactual language or not it is at least in some cases possible to *measure* an average causal effect more or less directly. This is fortunate in a sense: recall that we are largely interested in questions of control. EoC is the relevant type of question here, since we would like to know whether making a certain change *in the future* will generate different outcomes.

2.3 Connections, lesioning and physical models

If we view agents as physical systems, one of the most intuitive ways to think of EoC is in terms of connections between components (or the removal of components or connections, i.e. lesioning).

[5] Indeed, as Dawid points out, there could in theory be any degree of correlation between the random component of Y_R and the equivalent in Y_L, so there is an infinity of possible positions for different statisticians to take.

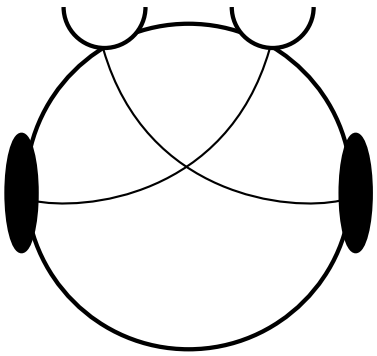


Figure 2.1: Top view of a phototactic Braitenberg vehicle. The semicircles represent light sensors at the front of the robot, and the elliptical shapes are wheels. The connection from the light sensor to the attached wheel is excitatory – i.e. the wheel will turn faster when more light is visible. Thus if a light is presented on the left hand side, the right hand wheel will turn faster than the left hand wheel, resulting in a left turn. Various "vehicles" of this sort were proposed by Valentino Braitenberg (Braitenberg, 1984) to demonstrate how simple connections from sensors to motors could produce seemingly purposeful, autonomous behaviour.

Slightly more generally, we can think of physical coupling strength, which may take different values (rather than simply connected or not connected). For example, our phototactic robots could be Braitenberg vehicles (Braitenberg, 1984) as illustrated in Figure 2.1. Since the left hand light sensor is connected to the right hand motor and *vice versa* the robot will turn towards any single light source presented.

Suppose we "lesioned" (i.e. removed) the connection from the left hand sensor – the right hand motor will no longer turn, and so the robot can only change direction towards the right. This gives a simple demonstration of the role of the lesioned connection, both in terms of its most direct effect (it allows light on the left to cause the right hand motor to turn) and more indirectly its connection to the overall behaviour (it allows light on the left to cause the agent as a whole to turn left).

We can also quantify the strength of connection: for example perhaps there is an approximately linear relation between the intensity of the light I and the rotation speed of the motor ω according to some coefficient a:

$$\omega = aI$$

Lesioning the entire connection corresponds to setting $a = 0$, but of course it could simply be adjusted, so there is a range of connection strengths.

The "cause" here is the light intensity I, and the effect is the motor speed ω. The above

equation does not make this clear however: it could obviously be re-arranged $I = \omega/a$. To be interpreted causally, there must surely be some kind of directionality: either I causes ω or ω causes I, but not both. We ought to adopt some different notation then to indicate what is the cause and what is the effect, for example:[6]

$$\omega \leftarrow aI \tag{2.1}$$

This can be thought of in terms of *atomic manipulations*. If we make a small change to I, so that the light intensity becomes $I + \Delta I$, then clearly the resulting speed changes by $a\Delta I$. On the other hand consider what happens if we alter the speed ω by some method other than changing the light intensity (this is what we mean by the manipulation being *atomic* – we suppose that at least in principle we can enter the system and change a variable by some means other than altering its causal antecedents within the system). For example most motors can be pushed externally to go faster than they would under their own power, but clearly doing this would not alter the light intensity.

Thus equation 2.1 is a simple causal model. From it, we can make several predictions, which we could group into two categories:

- Effect of manipulating the *inputs and outputs of the system*:
  - Altering I will alter ω proportionally.
  - Altering ω will not change I.
- Effects of changing the *system itself*:
  - Modifying a alters the strength of the influence of I on ω.
  - As a special case, setting $a = 0$ will result in $\omega = 0$ at all times.

[6]This general point is often made in the literature on causation, but the specific notation used for disambiguation varies. Pearl (2009, ch. 5, pp. 159-60) addresses the point but continues to use the equality (=) sign in his structural equation models, taking this caveat as read. Others adopt various notations, c= for example. I use $\leftarrow$ which is familiar to computer scientists as representing "assignment" – the operation which places the quantity on the right into the variable on the left, which is distinct from mathematical equality. For example, $x \leftarrow x + 1$ can be naturally read as an assignment that increments the value of x, whereas $x = x + 1$ is an equation with no solutions. As well as making it clear that $a \leftarrow b$ means changes to b affect a but not vice versa, this notation imports another useful implication from its interpretation as assignment, *viz.* a global ordering on the variables. That is, the net effect of combining $a \leftarrow b$, $c \leftarrow a$ and $b \leftarrow c$ clearly depends on the order in which the operations are performed (as it would if they were a listing of a computer program). Likewise a set of variables has a "causal" ordering – the order imposed when we place causes before their effects.

The symbols ω and I are viewed as variables: values that change during the normal course of operation of the system, whereas a is a parameter – describing the system itself – which is usually fixed. We could make this more explicit in the model by writing it as:

$$\omega \leftarrow f(I; a)$$

The symbols after the semicolon are parameters – these describe the *system itself* rather than variables *processed by the system*. These could come in various forms:

- Constants (a in the current example) being the most obvious kind.
- Changing but independent parameters: some parameters might change over time during the operation of a system, but independently of the variables that the system operates on.
- *Adaptive* parameters: those which change in response to changes in the system variables.

It is primarily the variables that are considered as potential causes and effects: the light intensity causes the motor rotation, but the parameter a is a measure of the *strength* of this causal influence, rather than a cause in and of itself. This jars with the fact that equation 2.1 shows no obvious difference between a and I – given that equation alone it would seem reasonable to call a a cause of ω just as much as I. The distinction between variables and parameters is for the most part down to semantics: it is up to us to judge what constitutes a variable and what constitutes a parameter, a judgement which comes down to what is an *input* or *output* of the system (i.e. a variable) and what is part of the *description* (a parameter) of the system. Yet it would also be perfectly reasonable according to a slightly different perspective to consider the connection strength a be a variable, particularly in the following example where a is modified to determine its causal role.

Of course a system can be composed of multiple variables – in a more complete model of the Braitenberg vehicle we would have at least two equations, one for the left wheel and one for the right. In such systems, as we have noted, it is natural to consider lesioning or direct intervention on either the system components or parameters to investigate causal relationships. Broadly speaking, this approach corresponds to what is known as "Mill's Method of Difference" after J. S. Mill[7]. That is, if we take a system, then vary some aspect (whilst keeping everything else as much as possible the same), then the effect of that intervention should tell us the causal role of that component which we varied.

Convenient as this is, it is known to have a number of problems, due to the complex relationships between input variables and output results that can easily occur in systems with multiple

[7]See e.g. Cartwright and Munro (2010) or Mill (1843, ch. VIII). Recall also the logic of the RCT discussed in section 2.1.

components. For example, take a simple model of the phototactic Braitenberg vehicle where the left sensor activates the right hand wheel and vice versa:

$$\begin{aligned} \omega_L &\leftarrow a_L I_R \\ \omega_R &\leftarrow a_R I_L \end{aligned}$$

Such as system with a light source at a given location can be simulated with a differential equation model.[8] This time, we treat the connection strengths a_L and a_R as variables (rather than parameters) which will be varied to investigate their causal effects.

To start with, set the connection strengths $a_L = a_R = 6$ which results in the robot following the path to the light shown in Figure 2.2. However we can significantly inhibit the left hand sensor to right hand motor connection by setting variable a_R to only 1 whilst keeping a_L the same. The result of this is also shown in Figure 2.2. Since the right hand motor turns a little slower in general, the robot veers slightly to the right of its target at first, but as it gets closer, this effect is naturally corrected by the feedback loop in the system. This is of course a useful effect – it means that in general we do not need the connection strengths a_R and a_L to take exactly the same value (which would be difficult to achieve in a real robot). But it also means that the causal significance of the inhibited connection could easily be underestimated – we know (by construction of the example) that both connections are important, but the system as a whole acts almost as if it is obscuring this fact from us when we attempt to study those connections by varying them. If we looked only for example at the final position of the robot relative to the light source, it would appear that inhibiting the left hand connection had almost no "causal effect", at least not until the inhibition was extremely severe (as we noted at the start of this section, total lesioning of the connection would of course mean that the right hand motor would not turn at all).

[8] Specifically, let the distance and direction to the light source relative to the current position of the robot be s and θ_s. The robot has a current heading θ_h. Let the overall forward speed of the robot be $v = (\omega_L + \omega_R)/2$ (assuming the wheels are of equal radius which we will call 1 in arbitrary units). The system can be described by the time differentials $\dot{s} = -v\cos(\theta_s - \theta_h)$, $\dot{\theta_s} = \omega_s = v\sin(\theta_s - \theta_h)/s$ and $\dot{\theta_h} = \omega_h = (\omega_L - \omega_R)/d$ where d is the diameter of the robot, dots represent differentiation with respect to time. Solutions to the initial value problem for these three equations give the trajectory the robot takes when the light source is placed at an initial location $(s(0), \theta_s(0))$ and the robot starts with heading $\theta_h(0)$. The values of ω_L and ω_R are obtained from the causal model given in the text which chosen parameters a_R, a_L. To get the light intensity values I_R, I_L, we model the total light intensity as decreasing according to an inverse square law and directional light sensors with maximum response at a given angle relative to the heading of the robot. The equation used in the current example is $I_i = \exp(-w(\theta_s - \theta_h - \theta_i)^2)/s^2$, $i \in L, R$ where w is a fixed shape parameter (value 6) and θ_i is the direction of maximum response for sensor i, namely $\theta_R = 0.7$ and $\theta_L = -0.7$ (radians).

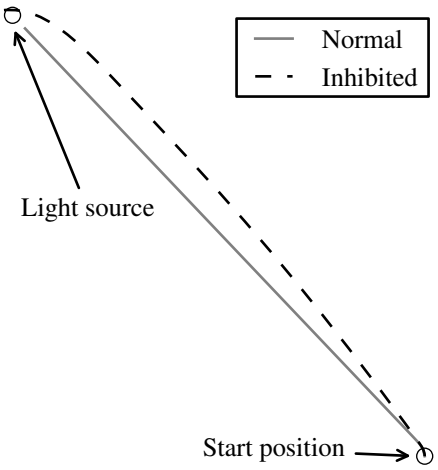


Figure 2.2: Inhibition experiment. The path traced by a phototactic vehicle from its start location to a light source is shown for both a standard connection between sensors and motors, and when the connection from the left hand sensor to the right hand motor is significantly inhibited. Even with inhibition, the behaviour is broadly the same, (the robot still navigates quite successfully to the light source), which would seem to underestimate the causal significance of the connection.

This shows how with only a fairly straightforward system we can get complex and non-linear input-output relationships that create problems for attempting to make causal inferences in a reductionist manner. In practice, systems can exhibit a variety of behaviours as a result of these complex relationships. A classic lesioning experiment in cats showed that while lesioning one brain area (the right occipito-temporal cortex) impairs vision in the left-hand visual field, subsequently removing another one (the left colliculus) can restore it (Sprague, 1966). Another example is that of gene "knock-out" experiments (Mitchell, 2009, ch. 4) – similar in principle to lesioning experiments, techniques for knocking out genes (i.e. replacing a particular chromosomal sequence with another, inactive one) have been developed in order to infer which bits of genetic material produce which phenotypic traits. However, there is so much robustness in the developmental processes that produce macroscopic traits from combinations of genetic sequences that such experiments have identified the influences of far fewer genes than was expected – much of the time, knocking out a gene just has no effect.[9]

[9]The proportion reported by Mitchell (2009, p. 67) is 30% of gene knock-outs have no effect. Of course it is difficult to say *why* such knock-outs have no effect – perhaps in some cases those genes really do nothing, that is to say, have no causal influence on phenotypic outcomes. However Mitchell and others argue that biological systems often exhibit considerable robustness to such interventions, which plausibly results from their having evolved to mitigate deleterious mutations that might occur naturally (Kitano, 2004; Edelman and Gally, 2001). This means even when a given gene has a role in producing a given trait under "normal conditions" (and thus can be appropriately considered causal), it might still be the case that knocking out that gene may not destroy the phenotypic trait, because some other (set of) gene(s)

To summarise this section: physical models can potentially be used to describe causal relations, but while we could use parameter values to measure "strength" of causal influences, we should be cautious as changes in parameter values will not always lead to linear changes to observable outcomes.

2.4 Probabilistic cause

So far the discussion of causation has been mostly limited to scenarios that are easily modelled as deterministic effects. In the Humean formulation, causes always precede their effects, giving a strong regularity: if A is always followed by B, then A is a cause of B. However many causal relationships do not fit easily within this structure, motivating the use of *probabilistic* definitions of causality.

The theories we will discuss in this section derive from the work of Hans Reichenbach (1891-1953), in particular his concepts of "screening-off" and the well known "principle of the common cause" which we will come to shortly (Reichenbach, 1956). Similar concepts were developed in the same era by Good (1961a,b) and Suppes (1970). Probabilistic causation is closely related, as we will see later in this section, to *causal Bayes nets*, and also to Granger causality (Granger, 1969; Holland, 1986; Seth, 2007) which will not be discussed in detail in this chapter, but appears later in this thesis.

As a start, we define probabilistic causation by the claim that *causes raise the probabilities of their effects*. Put symbolically, event c is a cause of e when:

$$P(e|c) > P(e|\neg c) \tag{2.2}$$

Where $P(e|c)$ means the conditional probability that event e (the effect) occurs given that the event c has already occurred. The event $\neg c$ is defined as the non-occurrence of event c.

Immediately we have solved one problem with the pure-regularity approach to causation: for example, smoking is thought to be a cause of lung cancer, but it is not the case that smoking invariably leads to lung cancer. This is not a problem for the probabilistic account provided that smoking *increases* the probability of lung cancer.

However this does not solve the problem of *spurious correlations*. Events might occur together without one being a cause of the other: in a deterministic setting, perhaps an incandescent bulb always produces light and heat at the same time, but one is not a cause of the other, they are both common effects of the electric current passing through the filament. Similarly, one might try to

"takes over" that same role.

"explain away" the increased probability of lung cancer in smokers by arguing that there might be a common cause, perhaps some genetic factor, which leads people both to smoke and gives a propensity to lung cancer.[10]

Let us modify the definition of probabilistic cause then, to try and avoid the problem of spurious correlations. We can do this by adding a *ceteris paribus* or *all else equal* clause, that is: *all else equal, causes raise the probabilities of their effects*. Symbolically, introduce an event k_i which represents the event that "all relevant background factors obtain a particular combination of values" (c.f. Cartwright, 2007a) where each combination of values for the background factors has been indexed by i, so:

$$\forall k_i : P(e|c, k_i) > P(e|\neg c, k_i) \tag{2.3}$$

is our new definition of probabilistic cause. In the case of smoking and lung cancer, this is akin to requiring that the probability of lung cancer is higher for smokers whenever any possible genetic common causes are the same. This appears to be a much more satisfactory definition, though it still has its problems. The most obvious being that it may be difficult to determine what the "relevant background factors" are, much less to ensure that they have been kept fixed. If we include too many background factors, for example each k_i is restricted to the exact circumstances of a particular individual, this definition would require that the probability of lung cancer would be increased for *every single individual* if they smoke. This raises first an epistemological problem, much like the "fundamental problem of causal inference" we saw earlier – there is a sample size of one for each k_i, and in each case we will only observe either c or $\neg c$. Moreover, even if we had some oracle which tells us the relevant probabilities, this overly restrictive choice of background factors would not really address the question we are interested in, which is whether smoking *generally* causes lung cancer, rather than whether it does so for every particular individual.[11]

2.4.1 Common causes

Thus to make further progress we ought to consider in more detail how to handle common causes. In this section we will look at Reichenbach's *principle of the common cause* (Reichenbach, 1956) and the related *causal Markov condition*. Intuitively, these principles encapsulate the view that although dependence of two variables clearly does not guarantee that one causes the other, there must be *some kind* of causal explanation. In other words, (true) dependencies do not simply happen "due to chance". Formally, the principle of the common cause (PCC) can be stated:

[10] See Pearl (2009, pp. 83-85) for a discussion of this scenario.

[11] See Cartwright (1994, pp. 55-58), for further discussion of this definition of probabilistic causation and the difficulty of choosing appropriate background factors.

(PCC) If X and Y are statistically dependent, then either X causes Y, Y causes X or there is a common cause Z of both X and Y such that X and Y are statistically independent given Z.

It is important in the formal definition to refer to "statistical dependence" rather than correlation. The latter may have a number of confusing interpretations, and moreover generally only refers to linear relationships. We will also treat the above definition in terms of *variables* X, Y, Z etc rather than individual events. A random variable is simply an assignment of particular events to values. Formally, define an event space Ω which contains events ω, a random variable X is simply a function of the event space which has a range known as its *support*, e.g. $X : \Omega \to \mathcal{X}$. Particular values taken by the random variable will be written in lower-case, and probabilities are assigned to values according to the probabilities of their underlying events. So the probability that X takes some value x is the probability of the preimage of x in X:

$$P(X = x) = P(\Omega_x) \text{ where } \Omega_x = \{\omega \in \Omega : X(\omega) = x\}$$

When we talk of X causing Y in the sense of variables then, we are not restricting ourselves to particular events like the cause and effect events c and e from the previous section, though the conceptions are broadly equivalent, since events can simply be thought of as "the event (or set of events) that occur(s) when variable X takes the value x".

Now we can formally write statistical independence of X and Y as

$$\forall x, y \in \mathcal{X} \times \mathcal{Y} \quad P(X = x, Y = y) = P(X = x)P(Y = y) \tag{2.4}$$

Or in less cumbersome notation simply

$$P(X, Y) = P(X)P(Y) \tag{2.5}$$

where it is implicit that the relation holds for all possible combinations of X, Y values. $P(X, Y)$ is called the *joint probability* of X and Y, and represents the probability of a particular combination of values occurring together, whereas $P(X)$ is the *marginal probability* of X represents the probability that X occurs irrespective of the value of Y. Probabilistic variables that satisfy equation 2.5 are said to be statistically independent, those that do not are statistically dependent. A shorthand notation for statistical independence is $X \perp\!\!\!\perp Y$, i.e.

$$X \perp\!\!\!\perp Y \Leftrightarrow P(X, Y) = P(X)P(Y)$$

For example, two separate fair coin tosses are statistically independent, because the probability of a heads on either coin (the marginal probability) is ½, and the probability of heads on both coins (the joint probability) is ¼ (i.e. the product of the marginal probabilities – clearly this holds for all combinations of heads and tails results, satisfying equation 2.5). On the other hand, suppose two archers stood next to each other shoot at two separate targets, and we model this as two random variables $H_i, i \in \{1, 2\}$ which take the value 1 if archer i hits the target and 0 otherwise. Each archer may have an average probability p_i that they will hit the target, and for the sake of argument let both probabilities by $1/2$ as before. In this case we would not expect the probability that both archers miss to be $1/4$ if some of their misses are due to random gusts of wind that affect both archers – if one archer misses due to a gust of wind, then the same gust of wind will affect the other archer. As a result the joint probability of both of them missing, $P(H_1 = 0, H_2 = 0)$, will be greater than the product of the marginal probabilities. This means that the variables H_1 and H_2 are not statistically independent, and so PCC says that either one causes the other or there is a common cause – in this case we are supposing that the wind is the common cause.

The PCC also makes a further claim in the case of common causes, namely that if X and Y are statistically dependent due to a common cause Z, then Z should "screen off" X from Y, that is to say, X and Y should be *conditionally independent* given Z, written $X \perp\!\!\!\perp Y|Z$. In terms of probabilities, conditional independence is defined as:

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Dividing the above by $P(Y|Z)$[12] and using the probabilistic identity

$$\frac{P(X, Y|Z)}{P(Y|Z)} \equiv P(X|Y, Z)$$

we see that conditional independence is equivalent to:

$$P(X|Y, Z) = P(X|Z)$$

This second formulation (though mathematically identical to conditional independence) captures what is meant by "screening off" – the conditional distribution of X is unaffected by the value of Y once Z is known, thus Z "screens off" X from Y when $X \perp\!\!\!\perp Y|Z$. Note from the above definition of conditional independence it is clear that conditional independence is symmetric in the sense that:

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow Y \perp\!\!\!\perp X|Z$$

[12] We have to assume here that $P(Y|Z)$ is always greater than zero.

so we could equally say that Z screens off Y from X whenever Z screens off X from Y.

In the case of the archers, say that we only look at cases where there is no wind, for example a third "wind variable" W takes the value 0. Now the probability of either archer missing the target will be lower than the average: say $P(H_i = 0|W = 0) = 1/4$. The PCC states that since the wind is held fixed (and assuming wind accounts for all the common causes of both variables), the probability of both of the archers hitting the target is just $(1/4)^2$ – i.e. $1/16$. That is, holding the common cause variable fixed should lead to statistical independence in the remaining system.

There is no "proof" of the PCC – is it not a mathematical theorem, rather, it captures a certain intuition about causation, as stated at the beginning of this section, that strong probabilistic relationships between events (beyond mere coincidence) must have some underlying causal explanation. There is a great deal of utility in this principle: supposing the two archers' shots are statistically related, what reason could there be for this other than some causal explanation? Nonetheless, there are a few caveats and "stock objections" to the general principle. A review can be found in Arntzenius (1992). Here I will describe two general classes of objection relating to ways in which variables might be probabilistically related in spite of no causal connection, thus violating the PCC: first, variables which are constrained by some non-causal requirement, and second variables which are correlated in time series despite not having a common cause. [13]

The first general class of objection then comes about where variables have some *non-causal constraint*. There are a few subtly different ways in which such a constraint might come about, but consider the following examples:

- A specific particle of an ideal gas x is moving around inside an enclosed space S. Define a binary partition of the enclosure as two spaces S_1 and S_2 such that S_1 and S_2 do not intersect. Then the joint probability of finding the particle x in both of the partitions, $P(x \in S_1, x \in S_2)$ is necessarily zero, but the marginal probabilities $P(x \in S_1)$ and $P(x \in S_2)$ are in general non-zero (if S_1 and S_2 both define volumes where it is possible to find x), thus the product of the marginal probabilities is strictly positive and does not equal the joint probability of zero.

- A resistor with Ohmic resistance R has a current I flowing through it, where I is drawn from some probability distribution $f_I(I)$, for the sake of argument a uniform distribution over the interval $(0, 1)$. Ohm's law then tells us that the voltage across the resistor V is distributed as $f_V(V)$ which in this case is uniform over the interval $(0, R)$, since $V = IR$. We can again construct two marginal probabilities the product of which does not equal the joint probability – for example with the intervals $I_0 = (0, \frac{1}{2})$, $V_0 = (\frac{R}{2}, R)$, we have

[13] A third category that is not discussed here is non-local correlations in quantum systems (Hitchcock, 2012).

$P(I \in I_0) = P(V \in V_0) = \frac{1}{2}$, so the product of marginal probabilities is $\frac{1}{4}$. But we know from Ohm's law that the current and voltage cannot be in I_0 and V_0 at the same time, thus the joint probability $P(I \in I_0, V \in V_0) = 0$.

- For a standard deck of playing cards, consider the random variables for suit S with possible values $\{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$ and color C which is drawn from $\{R, B\}$ (red and black). Now if we pick a card at random, we have $P(S = \heartsuit) = \frac{1}{4}$ and $P(C = R) = \frac{1}{2}$, thus the product of the marginal probabilities is $\frac{1}{8}$. However, the probability of picking a *red heart* is clearly equal to the probability of drawing a heart, thus $\frac{1}{4}$ (since all hearts are red) – thus again the joint probability is not equal to the product of marginal probabilities.

These examples all describe situations in which case we have a joint probability that is not generally equal to the product of the equivalent marginal probabilities, which the PCC would have us believe implies a causal relationship. But in fact the relationship is a result of external *non-causal* constraints on the behaviour of the systems.

In the first case, the constraint is due to the stipulation that S_1 does not intersect S_2, thus precluding the possibility of finding x in both regions at the same time. The second example invokes a physical law which we take to be inviolable ($V = IR$), and the third relies on the (again we take it inviolable) definition of a standard deck of playing cards.

One might object that these examples do in fact involve a common cause in the sense that they describe alternative *readouts* of a genuine common cause, comparable to recording the same sound with two microphones (a case of a common cause – the sound – creating statistically dependent readings on the two outputs). Similarly, perhaps the actual position of the particle in the first example is being read out by two different measures – is it in S_1 and is it in S_2, the resistor example might be seen as having a true microscopic state (the velocities and potential energies of the electrons involved), where voltage and current are two ways to measure this state, and the particular card picked (like the particular location of the particle) is the true common cause of the suit and colour of the card.

The first problem with such an approach is that the invoked common cause does not temporally precede the subsequent variables at all. If the particle is in location x_1 at time t_1, then it is also in, or not in, space S_1 at that very same time, which violates the intuition that causes should precede their effects. One might counter that we can jettison the temporal precedence requirement but maintain the PCC. However, this could lead to an infinite regress – if we have chosen the true common cause as being the particular location of the particle (or the collective microscopic states of some electrons, or the particular card chosen), and we call that common cause event X, then we can also arbitrarily define the microscopic event $\neg X$ as the event that happens whenever X does

not – and a statistical dependence obtains in the same was as it did for $x \in S_1$ and $x \in S_2$ in the first example. Thus we require a micro-microscopic event X' that is the common cause of X and $\neg X$, *ad infinitum*.

The important point about these examples is that the statistical dependence is occurring due to the *way in which the events are defined* rather than as a result of reading out the same event with multiple instruments. Note for example that the second example should be taken in the sense that we know the exact voltage and current, and thus their relation is due to the nature of currents and voltages on a resistor. If instead we considered an experiment where we use two pieces of laboratory equipment to make separate measurements of current and voltage, then the common cause description is appropriate: the states of the measuring devices are causal descendents of the effects of currents and voltages. [14]

A second common category of objection relates to time series analysis. This is of much greater interest here since much of this thesis will be concerned with examining time series. The objection, most notably raised by Sober (1984, 2001), is that there are many pairs of time series that have similar rising or falling trends but which have no common cause. The prototypical example being British bread prices and Venetian sea levels – both of these have risen consistently over time yet clearly neither one causes the other nor is there a common cause. However, this objection confuses an *association* in some data with a genuine probabilistic relationship. For example, suppose bread costs 10p in year 1, and 11p in year 2, and the sea level goes from 0cm to 1cm over the two years (relative to its position in year 1). The fact that we have only two data points is unimportant (it may be a poor sample, but the same logic would apply if hundreds of data points were collected). Sober appears to regard the probabilities thus:

$$\begin{aligned} P(\text{Bread price in year } 1 = 10) &= \tfrac{1}{2} \\ P(\text{Sea level in year } 1 = 0) &= \tfrac{1}{2} \\ P(\text{Bread price in year } 1 = 10 \text{ and sea level in year } 1 = 0) &= \tfrac{1}{2} \end{aligned}$$

The logic being that ½ is the "proportion of the time" that the given values are observed in the data. Clearly if we allow this then we have statistical dependence and thus the PCC requires that we find a common cause.

However, there is no justification for regarding these probabilities as correct: the probability that bread costs 10p in year 1 has been estimated by comparison to what it costs in *both* years 1

[14] For further discussion of this objection see Arntzenius (1992). In a similar vein, Lewis (1986) requires that "events" not be defined disjunctively – for example the event *red* could be viewed as the disjunction of the events *hearts* and *diamonds* (disjunction meaning "or", as in "red is hearts *or* diamonds") – in a sense this is why the spurious correlation arises.

and 2. Such an estimate relies on an assumption referred to as *ergodicity* – a system is said to be *ergodic* if the proportion of time it spends in a particular region of its state space (over a relatively speaking long period of time) approaches the probability of that system being in that region of state space.

The example of the bread prices and sea levels is a system which does not appear to be ergodic – it belongs to one large class of non-ergodic systems, namely *non-stationary* systems, where there is a long term trend in the data.

The stationarity problem is discussed by Hoover (2003), but we should remark that the more fundamental problem is not the absence of stationarity, but the failure of the assumption of ergodicity (and non-stationarity implies non-ergodicity). It is due to non-ergodicity that we cannot esimate the probability distribution of the price of bread in a given year from the historical record, because *for each given year* we have a sample size of one (no matter how many years in total we record). But the same problem can arise for somewhat different reasons – there are other ways in which a system may be non-ergodic (while still potentially being stationary), a notable class of such systems being those with uniform repeating cycles, which will be seen several times in the course of this thesis (see for example section 4.5.1).

A common way to deal with the problem, as noted by Hoover (2003), is to find a related time series that *is* ergodic. For example, we could differentiate the two time series and look at how much they change year-on-year: in other words, subtract the constant rising trend from each series and look at the dependence between the "remainders". Provided we have a good justification for our probability estimates in the first place (in the case of time series, the justification is the ergodicity of the underlying system), we do not seem to run up against this particular issue with the PCC.

2.5 Causal Bayes nets

Notwithstanding these objections, the principle of the common cause appears to have at least some validity, but it is limited to (at most) "three-variable" systems. The recent development of *causal Bayes nets* can be thought of as a generalisation of Reichenbach's logic to more complex systems of causal interactions, involving potentially arbitrary numbers of causal variables. Causal Bayes nets are related to the *structural equation models* introduced by Sewell Wright (1889-1988) (Wright, 1921) – Wright realised that the *structure* of causal relationships between several variables (the facts about what causes what) could be used to determine various statistical properties irrespective of the exact nature of that relationship. The modern formalisms, as we will see, are based on a very similar principle.

First, a note on etymology. The term "Bayes net" derives from Bayes' theorem, which is succinctly stated as the following relation between conditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem relates a conditional probability to its "inverse" – $P(A|B)$ to $P(B|A)$ in this case. It should not be confused with *Bayesianism*, which is a philosophy of statistical inference (Fienberg, 2006).[15] Bayes' theorem allows one to calculate a conditional probability based on either its inverse as above, or perhaps more directly from the definition of conditional probability:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

If A and B are statistically independent, this calculation becomes rather trivial, since statistical independence means $P(A,B) = P(A)P(B)$:

$$P(A|B) = \frac{P(A)P(B)}{P(B)}$$
$$P(A|B) = P(A)$$

Thus the independence relationship can be useful for simplifying the calculation of a conditional probability. For our purposes a Bayes net is a directed acyclic graph (DAG) that can be used to encode the statistical dependence relationships between any number of variables, by assigning each variable a vertex, and drawing an edge between a parent and a child vertex if the parent variable is not statistically independent of the child variable when conditioning on any other set of variables. This will become clearer in the following discussion.

However, a given set of statistical independences can generally be encoded using multiple different Bayes nets. A *causal* Bayes net is a Bayes net where not only are the statistical independences encoded, but arrows are drawn only from "causes to effects". That is, the causal ordering of the variables is respected. In ideal circumstances, this leads to only one possible "correct" Bayes net, which is inferred as the true set of causal relationships, though as we will see this is not always trivial.

The constructions of causal Bayes nets used by Spirtes et al. (2001) and Pearl (2009) differ in detail, but both begin by developing a more complete version of the PCC for arbitrary numbers of

[15] Bayes nets are frequently referred to as *Bayesian* net(work)s, which is confusing, since there is nothing about Bayes nets that necessitates Bayesianism (Murphy, 1998). A Google Ngram search reveals that *Bayesian network* seems to be a far more common term as of December 2013 `http://bit.ly/1bEWr4v`, I will stick with Bayes net which seems less misleading, and since the methodology I am proposing to use is not explicitly Bayesian.

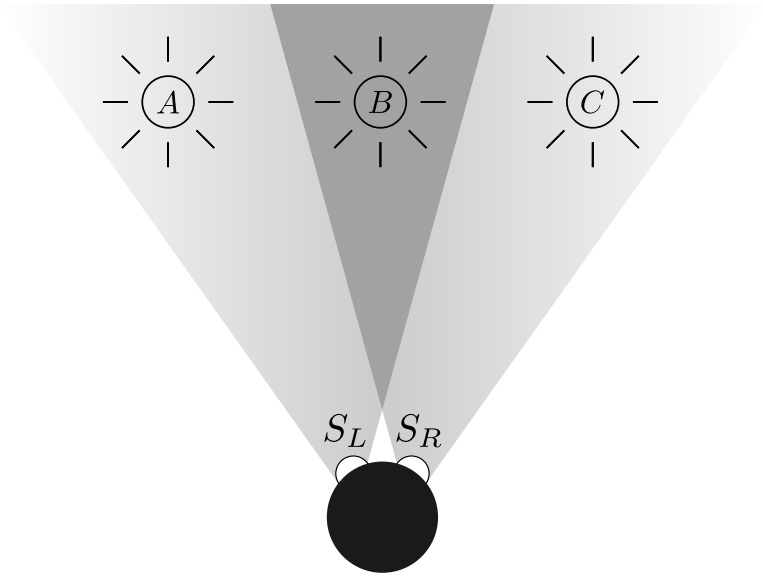


Figure 2.3: A phototactic robot with two light-sensors in an environment with three lamps.

variables, namely the *causal Markov condition*. To see how this works, we first need to construct the concept of a causal graph – a DAG where vertices represent variables and directed edges represent causal influences between variables. In the following discussion I will illustrate the causal Markov condition in a manner comparable to (though not identical to) the formulation used by Pearl (2009).

As an example, consider another phototactic robot, this time placed in an environment with three lamps in a row in front of the robot's starting position as illustrated in Figure 2.3. The three lamps can be given labels A, B and C.

Now let us imagine that the robot has two light sensors arranged at the front-left and front-right of its body. These have conical gain patterns such that the left sensor S_L can detect the light of either lamp A or lamp B, and the right sensor S_R can detect the light from either lamp B or lamp C. The arrangement is illustrated in Figure 2.3. Let us further assume that as before each light sensor drives the motor on the opposite side. So there is a torque for the left motor T_L driven by S_R and likewise a torque for the right motor T_R driven by S_L.

The arrangement suggests a physical model, defined similarly to the models we used in section 2.3. If we suppose we are considering a particular instantiation of the system, we can write the state of each of the components in lower-case (i.e. a is the state of lamp A in the current experiment). The physical model assigns values to each of the variables according to their causes. Assume that a function f is defined to perform the mapping for each variable:

$$s_l \leftarrow f_{s_l}(a, b)$$
$$s_r \leftarrow f_{s_r}(b, c)$$
$$t_l \leftarrow f_{t_l}(s_r)$$
$$t_r \leftarrow f_{t_r}(s_l)$$

Now, suppose that each of the variables in the physical model has some associated "noise" factor – that is, an argument is added to each of the functions representing some random influence. In the following we label the noise terms ε in the physical model:

$$\begin{array}{lcl} a \leftarrow \varepsilon_a & \qquad & s_l \leftarrow f_{s_l}(a, b, \varepsilon_{s_l}) \\ b \leftarrow \varepsilon_b & & s_r \leftarrow f_{s_r}(b, c, \varepsilon_{s_r}) \\ c \leftarrow \varepsilon_c & & t_l \leftarrow f_{t_l}(s_r, \varepsilon_{t_r}) \\ & & t_r \leftarrow f_{t_r}(s_l, \varepsilon_{t_l}) \end{array}$$

The noise terms serve a number of different functions although they are modelled in the same way. Supposing we are modelling an experiment where each of the lamps is switched on independently and at random, then the noise terms we have just added for the lamps, $\varepsilon_a, \varepsilon_b, \varepsilon_c$ perhaps model some system which makes the random choice as to whether to turn each lamp on.

On the other hand, the noise terms added to the rest of the model appear to represent something quite different – the noise ε_{s_l} for example models the imperfections of the sensor system, perhaps due to electrical noise or any other random physical impediment that stops the production of a certain amount of light (represented by the lamp states a and b) being enough to determine exactly whether or not the sensor will become active. Likewise ε_{t_l} represents some collection of random factors which influence the torque applied to the left motor, apart from the deterministic influence of the right hand sensor.

We can equally model the system using a distribution over random variables. For example, we have noted that the exact value of a is not fully determined by the physical model – there is a random influence. Thus we can designate a random variable A and give it a probability distribution – that is to say we include in the model the probability that A takes any given allowable value a, written $P(A = a)$.[16] Generally we denote the distribution as simply $P(A)$. Likewise, we have

[16] We will assume in this example that all probability distributions are over discrete variables. If A is a continuous valued variable then $P(A = a)$ gives a probability *density*. The distinction is not important for the present argument – we will stick to discrete valued variables for the sake of simplicity.

marginal distributions for all the variables: $P(B)$, $P(C)$, $P(S_L)$ and so on. However, the complete model is given by the *joint* distribution over all the variables:

$$P(A, B, C, S_L, S_R, T_L, T_R)$$

More generally, assume that there is a variable set $\mathbf{V}$ – the probability model is a joint distribution $P(\mathbf{V})$. This joint probability model encapsulates the probabilities of all the possible states of the entire system. Notice that if the physical model is fully specified (i.e. if we knew all the functional relationships f and the distributions that all the noise terms ε are taken from), then the probability model can clearly also be exactly obtained.

The causal Markov condition (CMC) tells us that certain conditional independence relationships must exist in the *probability model* (that is, the joint probability distribution) given only some limited information about the *form* (rather than the exact specification of) the *physical model* (that is, the set of equations that describe the behaviour of the system). By *form* I mean that we can see the causes of a variable looking at the right hand side of the assignment for that variable in the physical model equations. Using the current example, we can see that s_l is a function of a, b and ε_{s_l} so we call a, b and ε_{s_l} the (direct) causes of s_l. We know this even if we do not know the exact definition of the function f_{s_l} (nor do we know what value parameters to f_{s_l}, if there are any, might take – note that in the above model I have neglected to include any parameters in the physical model, since the CMC only requires information about the variables).

We begin by drawing a DAG G where the vertices of the graph are the variables from the probability model, and an edge exists from X to Y (pointing at Y) if and only if x is a direct cause of y in the physical model (i.e. we have a rule like $y \leftarrow f_y(\ldots, x, \ldots)$ in the physical model). Thus each edge in the graph represents a causal influence – the edge from A to S_L for example shows that lamp A is a cause for the left sensor. For the phototactic robot, the DAG is shown in Figure 2.4.[17]

There is a significance to the use of a *directed* and *acyclic* graph in this context. The direction of the arrows show the direction of causes to effects, and the lack of cycles encodes the assumption that it is not possible for something to "cause itself". This is perhaps natural in the phototactic robot example described here, but one may question how it could apply to systems that appear to mutually interact over continuous time – for example the trajectories of the earth and the moon through space are (presumably) "caused" by (among other things) the mutual influence of the gravity of each body on the other. This might lead us to imagine either an undirected connection between the bodies, or two directed connections (going each way and creating a cycle). However,

[17] Note that the random variables ε are not included in the DAG.

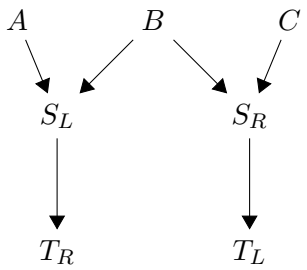


Figure 2.4: Causal DAG showing the influences between the components of the phototactic robot model.

if we view these interactions as causal, it still makes sense to refer in any given instance to one cause and one effect, without a loop that would seem to create an infinite regress. The typical solution to scenarios of constant mutual interaction is to "unroll" the causal DAG through time, and assign different nodes for different time points, with all directed connections pointing forwards in time (again, this fits a natural intuition that causes precede their effects). This approach will be seen several times later in this thesis, beggining in chapter 4.

The CMC, as formulated by Pearl (2009, p. 30 – Theorem 1.4.1), states that a given variable in the probabilistic model will be conditionally independent of its non-descendants[18] given its parents in the DAG, provided that all the error terms in the physical model are statistically independent of each other. We can write this more formally as:

$$\begin{aligned} &\forall(M, G = Cause(M), P = Prob(M), X \in V(G)) \\ &\qquad Markov(M) \implies X \perp\!\!\!\perp_P Nd(X;G))|Pa(X;G) \end{aligned} \tag{2.6}$$

Where M is any physical model, $Cause(M)$ is the causal DAG associated with M as described above, $Prob(M)$ is the probabilistic model associated with M, $V(G)$ is the set of vertices of DAG G, $Nd(X;G)$ is the set of non-descendants of X in G and $Pa(X;G)$ is the set of parents of X in G. The condition $Markov(M)$ is the requirement that all noise terms in M are statistically independent of each other – when this is satisfied we call M a *Markovian* model. The notation $\perp\!\!\!\perp_P$ is used to mean "statistical independence according to the probability model P". That is, if P is a joint distribution, then $X \perp\!\!\!\perp_P Y$ means we can derive $P(X,Y) = P(X)P(Y)$ from the

[18]Descendants of a vertex X are those vertices which can be reached by following directed edges from X without passing in the reverse direction along an edge. The set of non-descendant vertices of a vertex X is the complement of the set of descendants of X minus X itself. Note that X is not normally regarded as a descendant of itself, and thus the complement of the descendants of X would include X. The term "non-descendants" is thus perhaps somewhat confusing, but it is the term used by Pearl and others. It is clearly intended that a vertex should not be considered a non-descendent of itself, since if it were, this formulation of the CMC would be trivially false – a variable is generally not statistically independent of itself, irrespective of conditioning.

distribution[19].

The CMC is a generalisation of the PCC in at least as much as it makes similar predictions in the three-variable common cause scenario. Recall the example of two archers and the variables representing whether they hit their targets – H_1 and H_2 – with a common cause being the wind W. The PCC says that any dependence of H_1 and H_2 should be screened off by W – $H_1 \perp\!\!\!\perp H_2|W$. To apply the CMC to the same example, we can construct a model with w as a cause of both h_1 and h_2:

$$
\begin{aligned}
w &\leftarrow \varepsilon_w \\
h_1 &\leftarrow f_{h_1}(w, \varepsilon_{h_1}) \\
h_2 &\leftarrow f_{h_2}(w, \varepsilon_{h_2})
\end{aligned}
$$

Assuming the model is Markovian, the CMC predicts directly that $H_1 \perp\!\!\!\perp \{H_2, W\}|W$ – since $\{H_2, W\}$ are the non-descendants of H_1 and W is the only parent of H_1. This can be shown to also mean that $H_1 \perp\!\!\!\perp H_2|W$ – i.e. the CMC predicts the same conditional independence as the PCC:

$$
\begin{aligned}
H_1 \perp\!\!\!\perp \{H_2, W\}|W \Leftrightarrow \quad & P(H_1, H_2, W|W) = P(H_1|W)P(H_2, W|W) \\
\text{Using } P(A, B|C) \equiv P(A|B, C)P(B|C) : & \\
\Leftrightarrow \quad & P(H_1, H_2|W, W)P(W|W) = P(H_1|W)P(H_2|W, W)P(W|W) \\
\text{Using } P(A|A) \equiv 1 \text{ and } P(A|B, B) \equiv P(A|B) : & \\
\Leftrightarrow \quad & P(H_1, H_2|W) = P(H_1|W)P(H_2|W) \\
\Leftrightarrow \quad & H_1 \perp\!\!\!\perp H_2|W
\end{aligned}
$$

The CMC as I have formulated it here (following Pearl) is in fact very distinct from the PCC, because the CMC is a mathematical theorem and the PCC is not. In fact, the significant content of the CMC is not really in the mathematical implication shown in equation 2.6 – read literally, this states that *if* the physical model is Markovian, then any variable is conditionally independent of its non-descendants given its parents. The fuller meaning of the CMC (as commonly understood) is that all models *should be* Markovian. To Pearl (2009, p. 30), this reflects two points – first that a good model incorporates all common causes of any variables in the model (so for any two variables X and Y on a graph, all the common causes of X and Y should also be on the graph). This first requirement is sometimes called the *causal sufficiency* of a model. The second element of the CMC, and this is the real relation to the PCC, is the belief that if we do incorporate all common causes then it will be the case that the model is Markovian. In the archer example, we

[19] Just $\perp\!\!\!\perp$ without the subscript is used when the joint distribution referred to is clear.

do not believe there is a way that H_1 and H_2 could be dependent except that one causes the other or there is common cause (that is the PCC). Thus if in our model we do not have one causing the other, and we have also included the common cause as a separate variable w, the remaining random elements ε_{h_1} and ε_{h_2} *must* be independent of each other so as not to add a statistical dependence not accounted for by the common cause.

The consequence of this is that we can confidently argue that one cannot accept the PCC applies in some scenario and not also accept the CMC (in its full interpretation) – the latter only extends the former by rather incontrovertible logic. However, it also means that the CMC is subject to very similar limitations to the PCC. In the card selection example, there is no common cause of colour and suit, yet they are statistically dependent. Assuming we do not permit the actual card selected to be a prior event in the model (because it does not happen prior to the selection of the card's colour and suit), the physical model would simply be:

$$colour \leftarrow \varepsilon_{colour}$$
$$suit \leftarrow \varepsilon_{suit}$$

The associated DAG of course has only two vertices with no edges. If the model were Markovian, equation 2.6 would predict $Colour \perp\!\!\!\perp Suit$ (unconditionally, since there are no parent vertices). Of course the model is not Markovian (ε_{colour} and ε_{suit} are not independent), so it is not the CMC-as-theorem in equation 2.6 that is incorrect, rather it is the CMC-as-intuition that it should be possible to construct a Markovian physical model including all common causes that fails.

2.5.1 The CMC and d-separation

There is a second, somewhat more practical way to read the CMC. We have already seen for example that in the case where W was a common cause of H_1 and H_2, our current form of the CMC says that $H_1 \perp\!\!\!\perp \{H_2, W\}|W$ directly, but from this we can also derive $H_1 \perp\!\!\!\perp H_2|W$. In general, the CMC can be rewritten in a way that allows us to directly read off all the conditional independence relationships that exist in a graph of a Markovian model. This is a ternary relation known as d-separation, which describes all the conditional independences encoded in a graphical model (Geiger et al., 1990). In terms of d-separation, the CMC can be rewritten as follows:

> (CMC) Given a Markovian model M with its associated graph G and probability model P, if a set of variables $\mathbf{X}$ is d-separated in G from some other set $\mathbf{Y}$ by a third set $\mathbf{Z}$ (written $\mathbf{X} \perp_G \mathbf{Y}|\mathbf{Z}$, call $\mathbf{Z}$ the "conditioning set"), then $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$.

$$\mathbf{X} \perp_G \mathbf{Y}|\mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y}|\mathbf{Z} \tag{2.7}$$

Whether or not d-separation holds for a given choice of variables is a matter of whether there are paths between the two variables using the edges in the graph, and if so, whether members of the conditioning set appear along the paths. To begin with, consider the case of d-separation of two variables (rather than two sets of variables). By definition variable X is d-separated from variable Y by set $\mathbf{Z}$ – $X \perp Y|\mathbf{Z}$ – if any path from X to Y in the graphs is blocked by some member $z \in \mathbf{Z}$, where a path is "blocked" if it satisfies at least one of these conditions:

1. The path contains a chain $A \to B \to C$ or a fork $A \leftarrow B \to C$ with the "separating variable" B being an element of $\mathbf{Z}$.

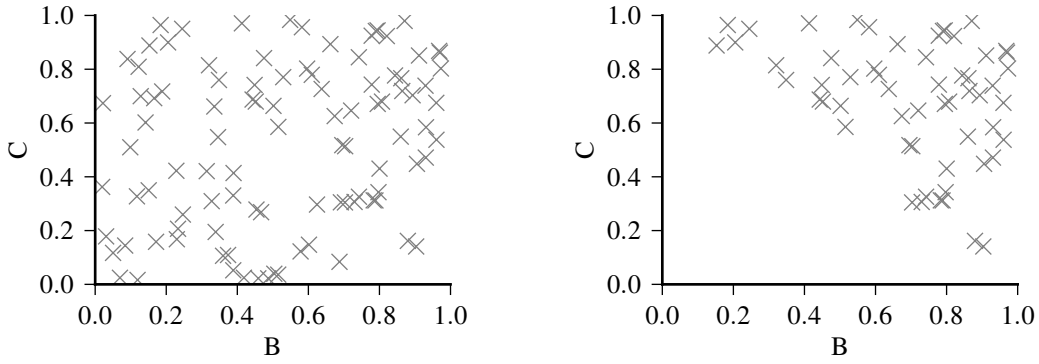
2. The path contains a collider $A \to B \leftarrow C$ where neither B nor any descendent of B is an element of $\mathbf{Z}$.

This allows us to arrive at $H_1 \perp\!\!\!\perp H_2|W$ in the archery example directly, because the only path from H_1 to H_2 is $H_1 \leftarrow W \to H_2$. Since W is in the middle of a "fork" on the only path from H_1 to H_2, condition 1 above tells us that $H_1 \perp H_2|W$ and this new version of the CMC gives the conditional independence directly.

The utility of this formulation is more obvious with more complex models: looking at Figure 2.4 again, there are a variety of conditional independence relationships we can expect by d-separation. For example, $S_L \perp S_R|B$ by the path $S_L \leftarrow B \to S_R$, and $C \perp T_L|S_R$ because the only path is $C \to S_R \to T_L$. These all rely on condition 1 of the d-separation rules above – this condition can be thought of as requiring that to d-separate two variables, we must "block" all paths that represent one of the variables causing the other (by conditioning on some mediating component), or those variables having a common cause (by conditioning on the common cause).

The second condition is more subtle. For the phototactic robot, we clearly expect that $B \perp\!\!\!\perp C$, since we have decided that the lamps B and C will be activated independently. We have $B \perp C$ (equivalent to $B \perp C|\emptyset$ – B and C are d-separated by the empty set) because the only path is $B \to S_R \leftarrow C$ – we have a "collider" and (importantly) the vertex in the middle in not a member of the conditioning set, thus satisfying condition 2 of the d-separation criteria. However, B and C are *not* d-separated by S_L, i.e. if we include S_L in the conditioning set, then neither condition for d-separation is met for the path $B \to S_L \leftarrow C$ and we have $\neg(B \perp C|S_L)$.

In this way, the d-separation requirement guards against "selection bias". The reasoning is illustrated in Figure 2.5. Suppose for a moment that the two lamps are switched on to an intensity level between 0 and 1, chosen independently at random for B and C. Then a scatter plot of 100



(a) No selection

(b) $B + C > 1$

Figure 2.5: Illustration of selection bias. If we condition on a function of two random variables (e.g. by considering only those cases where the sum of two variables B and C is greater than 1), then we may "introduce" a statistical dependence between the variables.

experiments with the two lamps looks like Figure 2.5a – there is clearly no dependence between the two variables. However, if the sensor S_R switches on, say, just when the sum of the light intensity from B and C is greater than 1, then if we only look at cases where the sensor is switched on (i.e. we look at the distribution $P(B, C|S_R = \text{on})$), we see a clear negative correlation between B and C (Figure 2.5b). In other words, if we know that the right hand sensor is switched on, then the light intensity of lamp B is predictive of the intensity of lamp C in the sense that the value of C must be at least $1 - B$. The second criteria of the d-separation definition ensures that we do not predict a statistical independence via CMC when we are conditioning on a "common descendent" such as the sensor state.

The rules of d-separation are straightforwardly extended to apply to entire sets of variables. We say that sets $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ if for every pair of variables from $\mathbf{X}$ and $\mathbf{Y}$ the d-separation criteria hold, i.e.

$$\mathbf{X} \perp \mathbf{Y}|\mathbf{Z} := \bigwedge_{X_i \in \mathbf{X}, Y_j \in \mathbf{Y}} (X_i \perp Y_j|\mathbf{Z})$$

Chapter 4 returns to the topic of d-separation and the CMC. To summarise this section, the succinct way to think of the CMC is as the implication in equation 2.7, which can be read as saying that (for Markovian systems), two variables must be independent when one conditions all common and mediating causes, provided one does not introduce "spurious" (non-causal) dependencies via selection bias. We have not, so far, discussed in much detail how this can be put to practical use in statistics – i.e. what can we tell about causes by looking at empirically collected data. This will be addressed later in this chapter. I will now turn to a rather critical point regarding causal Bayes

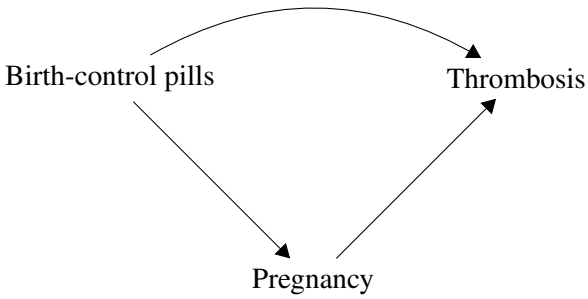


Figure 2.6: Graphical representation of the causal relationships between birth-control pills, pregnancy and thrombosis, following Hesslow (1976).

nets, namely that the CMC alone is not usually taken as the only assumption. In most applications, a complementary principle known as *stability* is also required.

2.5.2 Stability and faithfulness

Stability and faithfulness are the terms used by Pearl (2009) and Spirtes et al. (2001) respectively to refer to (roughly) the same concept – namely the "complement" of the CMC:[20]

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} \implies \mathbf{X} \perp \mathbf{Y}|\mathbf{Z} \tag{2.8}$$

That is, where the CMC says that d-separation in the graph implies conditional independence, stability states that if a conditional independence holds, then the associated d-separation criteria should be met. The peculiar thing about stability as compared to the CMC is that it is trivially false, even for Markovian systems. An oft-cited counter-example to stability derives from a scenario suggested by Hesslow (1976) is the effect of birth-control pills on thrombosis. Suppose that it is the case that taking birth-control pills increases the probability of thrombosis, yet it is also the case that birth-control pills reduce the probability of pregnancy, which is itself a potential cause of thrombosis. We can picture this as the causal DAG in Figure 2.6.

It is at least possible to conceive of a parametrisation of the system such that the probability of thrombosis would be the same irrespective of whether a person has taken birth-control pills. That is, we can have $P(\text{Thrombosis}|\text{Birth-control pills}) = P(\text{Thrombosis})$, or:

$$\text{Birth-control pills} \perp\!\!\!\perp \text{Thrombosis}$$

If stability is true, then it follows from this that taking birth-control pills is d-separated from

[20] Both Pearl (2009) and Spirtes et al. (2001) regard stability/faithfulness as essentially a complement of the CMC. But recall that they define the CMC in slightly different ways and I have been discussing the CMC primarily in the form used by Pearl, therefore I will refer to the complementary principle as *stability*.

thrombosis by the empty set. However according to our current model they clearly are not – in Figure 2.6 there are two paths from birth-control pills to thrombosis that do not fulfill the d-separation conditions with respect to the empty set. The problem is that the two causal chains "cancel out", giving no net effect. We therefore have a system which has a conditional independence that does not correspond to a d-separation in the causal DAG.

Note that it is quite possible for such a system to be Markovian. Supposing we have three real-valued variables x, y and z with independent Gaussian noise on each:

$$x \leftarrow \varepsilon_x \tag{2.9}$$

$$y \leftarrow \alpha x + \varepsilon_y \tag{2.10}$$

$$z \leftarrow \beta x + \gamma y + \varepsilon_z \tag{2.11}$$

If we set $\gamma = -\beta/\alpha$ then the net result is that $z = (-\beta/\alpha)\varepsilon_y + \varepsilon_z$. Since the system is Markovian, ε_y and ε_z are statistically independent of ε_x so we see that $Z \perp\!\!\!\perp X$ even though there is a path from X to Z since x appears on the RHS of the causal rule for z (this example is conceptually similar to the birth-control example).

In these examples there is no problem with the CMC, which recall only requires that any d-separations that appear in the graph have a matching conditional independence in the probability model – X and Z for example are not d-separated, and the CMC does not say anything about whether $X \perp\!\!\!\perp Z$. It is stability that says that in the absence of d-separation there *must be* a corresponding statistical independence (this being the contrapositive of equation 2.8).

Assuming or claiming that stability holds (let us call this the *stability assumption*) is therefore very different to claiming that the CMC applies. The two often appear side by side, and both are important parts of the frameworks used by Pearl (2009) and Spirtes et al. (2001), yet they are clearly logically distinct. Recall that the CMC effectively states that

d-separation implies conditional independence

whereas the stability assumption can be seen as a claim or requirement that

conditional independence implies d-separation.

These are logically unrelated, in the same sense that the statement

all presidents of the USA are persons born in the USA

is unrelated to the statement

all persons born in the USA are presidents of the USA

Why, then, should we be interested in stability, when it seems clear from the outset that it is not nearly as reliable an intuition as the CMC, and nor is it in any sense logically related to the CMC? Its primary role is to disambiguate possible causal graphs that can be inferred from a given set of conditional independences. I have presented the theory of causal Bayes nets above in terms of claims we can make about probability distributions assuming that we know the causal model or DAG. However, the end goal is in fact to be able to *infer* facts about the causal relationships in the world assuming that we have access to probability distributions (or at least approximate estimates probability distributions taken from data collected in the "real-world").

As a rudimentary example, suppose that we find that a robot's heading appears to be statistically independent of any light sources. Intuitively, the light source is not a cause of the robot's heading, but the CMC does not allow us to infer this directly – since it does not say *anything* about what we should infer about the causal graph when we see a statistical independence. It would still be consistent with the CMC if the light source was a cause of the heading even though the two are statistically independent. The stability assumption says that such a statistical independence will only occur if the light source and the robot's heading are d-separated in the causal graph – and this allows us formally to make the apparently intuitive inference of no causal connection.

From this we see that stability has the benefit that it may drastically narrow down the inferences that we might make. Inferences are made from facts about probability distributions to compatible causal graphs – stability states that if a probabilistic independence is present, it must have an equivalent d-separation in the graph – $X \perp\!\!\!\perp Y|Z \implies X \perp Y|Z$. On the other hand, the CMC states that a given observed probabilistic dependence implies that the associated d-separation is not present, via the contrapositive of equation 2.7, i.e. $\neg(X \perp\!\!\!\perp Y|Z) \implies \neg(X \perp Y|Z)$. Requiring both conditions to be satisfied narrows down the inferred causal graphs, as illustrated in Figure 2.7. A significant part of the literature on causal Bayes nets is devoted to algorithms that take the CMC and stability as assumptions and derive possible causal graphs in this way.

However, this does not negate the drawback that we do not have much reason to think that stability is true in the general case. For the phototactic robot it might seem quite reasonable: suppose that the light source is truly a cause of the robots heading, how could the two possibly be statistically independent if that is the case? In the case of the birth-control pills and thrombosis, we had a situation where one causal pathway "cancels out" the other one. For the phototactic robot, we expect that if the position of the light affects the robots heading at all, it will cause it to move one way or the other, rather than (via one causal pathway) make it move left and (via another pathway) make it move right an exactly equal amount.

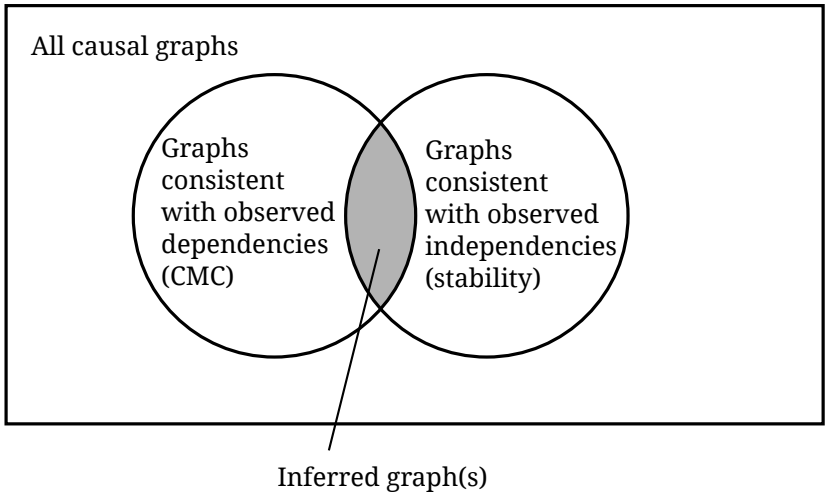


Figure 2.7: Using both the CMC and stability allows possible causal graphs to be narrowed down.

Problems with stability are often cited by critics of causal Bayes nets as a methodology. Cartwright (2007b, ch. 6) points to the birth-control-thrombosis example among others as problems with Bayes nets. Dawid (2009) notes that close mathematical relationships between variables (which we have already seen can be a problem for the CMC) can also create difficulties for the stability assumption. This is because we can easily find examples where stability contradicts itself if we require all conditional independences to be modelled by d-separation. One example given by Dawid is based on Boyle's law for isothermal (constant temperature) changes in pressure and volume of an ideal gas in a closed box:

$$PV = c \tag{2.12}$$

This states that pressure P and volume V are inversely proportional with proportionality constant c (this example has a similar form to the Ohm's law example in section 2.4.1, but here we are interested in how this relates to *stability* rather than the principle of the common cause.) If the pressure and volume are manipulated with a piston (carefully designed to avoid introducing temperature changes), which has state F, then we would find that

$$P \perp\!\!\!\perp F|V$$
$$V \perp\!\!\!\perp F|P$$

These independences occur because pressure and volume are exactly related according to Boyle's law, and thus once, for example, the volume is known, the pressure can be determined from equation 2.12. Since a given pressure only allows one possible volume, clearly the volume is conditionally independent of the piston state given the pressure: $V \perp\!\!\!\perp F|P$. A similar argument produces $P \perp\!\!\!\perp F|V$. However, since we also have $\neg(P \perp\!\!\!\perp V)$, the CMC requires that P and V

not be d-separated. There is no three-variable causal DAG for P, V and F that simultaneously satisfies this and the d-separations required by stability:

$$\neg(P \perp V)$$
$$P \perp F|V$$
$$V \perp F|P$$

That is, there is no causal DAG which fulfils both the CMC and the stability assumptions. Perhaps the most intuitive graph would be one in which F is a common cause of P and V ($P \leftarrow F \rightarrow V$), but that satisfies neither of the d-separations inferred from stability. Otherwise, we would have to make either P a cause of V or V a cause of P – but neither of these makes intuitive sense, and even if we do, it is impossible to simultaneously satisfy $P \perp F|V$ and $V \perp F|P$ with a single DAG.

Consider again the DAG for the phototactic robot in the three-lamp scenario (Figure 2.4). Imagine that the wiring between sensors and motors is made deterministic rather than probabilistic – so that the left motor T_L is active at exactly the same time as S_R, and inactive whenever S_R is inactive. In other words, remove the "noise" term from the $T_L \leftarrow$ rule. We would now have for example $S_R \perp\!\!\!\perp B|T_L$, since T_L tells us exactly whether S_R is active, irrespective of the otherwise causally-relevant state of B. Yet in the "correct" causal DAG in Figure 2.4 we can see that we do not have $B \perp S_R|T_L$ – i.e. the stability assumption fails in this scenario.

How do we know when stability is a reasonable assumption? Recently some have developed empirical methods for validating the assumption of stability from data (Zhang and Spirtes, 2008). More traditionally, some arguments have been given by proponents of causal Bayes nets as to why stability can be assumed in the absence of any particular evidence.

Pearl (2009) and Spirtes et al. (2001) give similar arguments, I will start with Pearl's. Pearl initially gives a weaker form of stability which is true. This can be phrased as follows: suppose that a given Markovian model M has a set of parameters Θ which determine the precise relationships between each of the variables. The exact distribution P generated by the model is thus also dependent on Θ. We know by the CMC-as-theorem that the distribution will have all the conditional independences required by d-separation, irrespective of Θ:

$$\forall\Theta, X \perp_{G(M)} Y|Z \implies X \perp\!\!\!\perp_P Y|Z$$

That is, d-separation is a *sufficient* condition for conditional independence according to the CMC. The weak version of stability (Pearl, 2009, Theorem 1.2.4, p. 18) says that there must be a choice of values for the parameters where d-separation is also a *necessary* condition for conditional independence:

$$\exists \Theta, X \perp\!\!\!\perp_P Y|Z \implies X \perp_{G(M)} Y|Z \tag{2.13}$$

That is, given that the parameters are chosen in the right way, all of the conditional independences in P will be "represented" by equivalent d-separation relationships in the graph. By contrast the strong (but false) stability assumption says that

$$\forall \Theta, X \perp\!\!\!\perp_P Y|Z \implies X \perp_{G(M)} Y|Z \tag{2.14}$$

(c.f. equation 2.8, where we neglect $\forall\Theta$.) By way of example, recall that in equations 2.9 to 2.11 we had a model where stability was violated only when the parameters $\Theta = \{\alpha, \beta, \gamma\}$ obey the exact relationship $\gamma = -\beta/\alpha$. When we use these parameters there is an independence in the probability distribution that is not represented in graph. However for any other choice of parameters, stability will hold. Pearl gives this explanation:

> The converse part of Theorem 1.2.4 [i.e. "weak" stability – equation 2.13] is in fact much stronger – the absence of d-separation implies dependence in *almost all* distributions compatible with G. The reason is that a precise tuning of parameters is required to generate independence along an unblocked path in the diagram, and such tuning is unlikely to occur in practise. (Pearl, 2009, p. 18)

Pearl's use of the language *almost all* (emphasis in original) and the phrase "unlikely to occur in practise" suggest that he is thinking of the parameters in a probabilistic or measure theoretic sense. The logical "for all parameters" ($\forall\Theta$ in equation 2.14) is replaced with the probabilistic "for almost all parameters". In our example with the parameters $\{\alpha, \beta, \gamma\}$, if these parameters are all allowed to take any real value, and we imagine that the "true" value is drawn from a continuous probability distribution, it follows that the probability of the exact relationship $\gamma = -\beta/\alpha$ is zero. All other choices of parameter values produce "stable distributions" (distributions which satisfy equation 2.8 with respect to the causal DAG), hence stable distributions occur with probability one – this seems to be what Pearl means by "almost all" distributions being stable.

Spirtes et al. (2001) make a very similar argument, saying that in many cases linear parameters "form a real space, and the set of points in that space that create vanishing partial correlations not implied by the Markov condition have Lebesgue measure zero".[21] Cartwright objects to the appeal

[21] The Lebesgue measure is the "volume" of a set in a measure space. A flat dartboard for example has a given area $A = \pi r^2$ which can be thought of as a Lebesgue measure of the set of all points on the dartboard. The bullseye is a smaller area $A_b = \pi r_b^2$. But the exact centre of a dartboard has Lebesgue measure zero since it is a point with no radius. An analogy to the Spirtes et al. argument would be that it is "impossible" for a dart to land on the exact centre of the dartboard, though it can obviously land somewhere inside the bullseye.

to Lebesgue measure, saying:

> I gather we are to conclude that it is unlikely that any causal system to which we consider applying our probabilistic methods will involve genuine causes that are not prima facie causes [prima facie causes are causes that result in statistical associations] as well.
>
> But this conclusion would follow only if there were some plausible way to connect a Lebesgue measure over a space of ordered n-tuples of real numbers with the way in which parameters are chosen or arise naturally for the causal systems that we will be studying. I have never seen such a connection proposed; that I think is because there is no possible plausible story to be told.
>
> Cartwright (2007b, p. 68)

Making a similar argument, Andersen (2013) points out that "[the rational numbers] are also measure 0 with respect to the real numbers, while irrational numbers are measure 1 ... However, this does not imply that rational numbers are unlikely to be encountered *simpliciter*".

Indeed why should we consider the parameters in terms of a measurable space at all? Notably, Cartwright's second paragraph appears to be worded carefully: there are numerous connections between parameter values and measurable spaces that have been proposed, namely any statistical methodology that comes under the rubric of *Bayesianism* assigns probability distributions (and thus measure spaces) to parameter values (it is unlikely that Cartwright is unaware of this). It seems plausible that Bayesian methodology informs Pearl's thinking in his previously quoted comment about "almost all" distributions. But perhaps it is telling that *if* Pearl and Spirtes et al. are thinking of any such appeal to Bayesianism, they leave it implicit. Bayesianism typically assigns measure spaces to parameters as a tool for *inferring what they are* (from the point of view of an agent who does not already know what they are), not for how they are "chosen or arise naturally," the words Cartwright uses. Indeed it is how parameter values arise in nature, not how they are inferred, that would be significant in determining whether or not the stability assumption holds.

Nonetheless, let us permit that it is unlikely that such "unstable" parameter values would occur at least in some cases. Perhaps the birth-control-thrombosis example seems implausible, since it does seem to be the case (per Pearl) that the net negative influence of the birth control pill on thrombosis (via reducing the probability of pregnancy) would be very unlikely to coincide exactly with the direct positive influence. But what allows us to view this as a typical example? In fact we have already seen one violation of stability that is not at all unlikely or implausible in the

inhibition experiment shown above in Figure 2.2. There we saw that changing the strength of an internal connection had almost no effect on the final position of the robot, in spite of the fact that surely the position of the robot is caused by the internal connections between sensors and motors.

To give perhaps a more intuitive example, your internal body temperature is more or less "decoupled" from the environment around you. It stays broadly fixed, though changes slightly in accordance with biological cycles and any illnesses you might have. But if you were to measure the temperature around you and your internal temperature you would find almost no statistical dependency – certainly none as strong as you would find between whether you have the flu or not and your temperature. But it is hardly reasonable to claim that the ambient temperature is not a cause of your body temperature – clearly it is, but your body has an internal homoeostatic mechanism which, as a rule, "functions" to cancel out any effect that the ambient temperature might have. Such regulation creates a circumstance much like the birth-control-thrombosis example – since the outside temperature is typically lower than your body temperature, it acts to cool you down, but your metabolism acts to heat you up, in a way that almost exactly cancels the external cooling. The net effect of zero does not seem at all implausible, since it is in fact indisputably part of typical operation of the of the system under study. Biological systems, by virtue of having evolved to maintain themselves against environmental perturbations, would seem likely to instantiate many stability/faithfulness violating systems (Andersen, 2013). Autonomous robots, inasmuch as they might self-regulate in a manner comparable to biological organisms, have a clear potential to produce similar problems.

There is a third principle that we have not discussed, somewhat related to stability, known as *minimality*. This says that when inferring a causal graph, one should not include any connections that are not required to make the graph compatible with the known probabilistic dependencies as per the CMC. Minimality requires that we do not have any "redundant" paths in our causal graph. There are cases where minimality alone (without stability) is sufficient to narrow down the causal graph to a single possible inference.

Thus minimality is another tool that (like stability) can exclude possible causal explanations. Pearl justifies this by appeal to Occam's razor – the principle that there is no good reason to add additional complexity where it is not needed. Thus minimality, like stability, can be used to "prune" unwanted connections from a causal graph. Some authors, such as Spohn (2001), consider minimality to be a more fundamental component of causal Bayes nets than stability. Dawid's Boyle's law example discussed above can equally be seen as a problem with minimality though – without requiring stability as such, we can still find *minimal* causal graphs for the variables P, V, and F – the problem is only that there are multiple such graphs which have the same number

of arrows and satisfy $\neg(P \perp V)$ as required by the CMC ($P \rightarrow F \rightarrow V$ and $V \rightarrow F \rightarrow P$ for example are equally minimal). Pearl gives a second presentation of stability (Pearl, 2009, p. 48) where it is seen as a sufficient condition for guaranteeing a unique minimal graph to exist. That is, if the observed dependencies have been generated from a stable distribution, then it follows fairly immediately that any arrows not required to produce known dependencies will be removed in the minimal graph, resulting in the same inference as the one obtained by using the stability-based methodology.

In this sense stability can also be seen as related to Occam's razor. The argument being that if there is no dependence between A and B, then we have no reason to think that there should be an arrow from A to B, therefore we do not add one (even though the CMC does not say that there cannot be a causal relationship in such a case). Thus we do not add additional "complexity" to the graph without justification. The problem remains that, as we have seen from the above examples, we do not in reality have any justification to say that there is not a causal connection either, and so we have no clear reason to think that this approach will lead to the inference of a graph representing the actual causal influences involved.

2.6 Summary

I have introduced graphical models of the sort advocated by Pearl (2009) as a primary tool for describing causal relationships. I have nonetheless taken care to highlight known issues surrounding this approach. Some of these are the subject of debates that are only tangentially relevant to the current work, but there are some points which will be critical for interpreting the remainder of this thesis.

First, we are concerned here with *control* and *general causal relationships*, rather than the specific questions about "what caused what" (recall the distinction between effects-of-causes and causes-of-effects). My position is that a reasonable framework for modelling such general relationships are what I have called *physical models* (conceptually very similar to what Pearl (2009) calls *structural equation models*.) It is important to bear in mind that models are, by their nature, *idealizations* – we assume that they are approximate and contingent on (often implicit) background factors being arranged in just the right way. As such, the models should be thought of convenient representations of the true causal relationships in the world.

Second, the *form* or *structure* (facts about what causes what) of such physical models can be captured in a causal DAG, which when interpreted as a Bayes net can inform us about independence properties we can expect from the probability distributions induced by a system. The PCC and CMC are useful concepts in this respect. This means that we can at least in some cases

make inferences by starting with some empirical data, determining the statistical dependencies, and working backwards to allowable causal DAGs.

Third, and significantly, this framework is far from perfect. There are clear cases where a naive interpretation of Bayes nets does not work. Problems with *stability*, far from being a "pathological" case, are extremely relevant. Systems that regulate their outputs via negative feedback (such as the Braitenberg vehicle or a thermostat) can lead to counter-intuitive results.

This framework will simplify later discussion of causality in this thesis, particularly in the contexts of transfer entropy and Granger causality.

Chapter 3

Stroboscopic transfer entropy

Synchronisation is a phenomenon of dynamical systems whereby oscillators, including chaotic oscillators, achieve a matched state when coupled (Pikovsky et al., 2001). Chaotic oscillators can match completely such that the trajectories of the two systems are identical (complete synchronisation) or only in terms of phase (phase synchronisation) (Rosenblum et al., 1996). By analogy to this distinction between the complete and phase-only information in a signal, this chapter proposes a method of calculating transfer entropy that incorporates only the phase information. We hypothesise that this method is of interest in the study of synchronising oscillators – relevant for this thesis because of our intention to use synchronisation as a method for generating robotic gaits – however, we find that this approach is still vulnerable to the non-separability problems relating Granger-type statistics to causal influences.

In relation to robotic gait generation, the model studied here is intended as an abstract model of an "active" dynamic walker (in contradistinction to a passive dynamic walker – see chapter 7). The system consists of active oscillators and passive resonant components. We consider the type of physical dynamics present in such a system, and perform "lesioning" experiments on the coupling influences to investigate the relationship between transfer entropy and causal influence.

This chapter is based on the paper published in the proceedings of the 2011 European Conference on Artificial Life (Thorniley, 2011).

3.1 Background

Central pattern generator (CPG) synchronisation is thought to underlie animal gaits (Collins and Stewart, 1993) and therefore the general phenomenon of synchronisation is of interest for this thesis. However, in a complex synchronising system it is often difficult to determine the nature of the interactions between oscillators. Ceguerra et al. (2011) approached this problem by using a

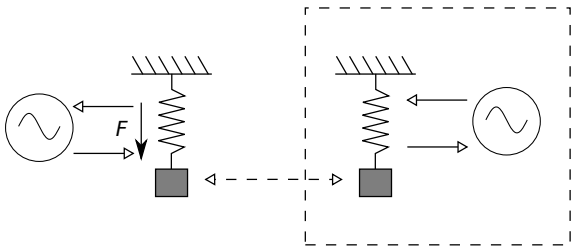


Figure 3.1: Illustration of coupled oscillator system. An electronic oscillator is attached to a mass-spring-damper system, providing force actuation and incorporating the resultant extension of the spring into the feedback path of the oscillator circuit. The model is extended by duplicating the system (dashed box) and coupling via the mechanical component.

form of transfer entropy to analyse the synchronisation process in networks of coupled oscillators, which was shown to be more effective than other methods.

This chapter investigates information transfer between oscillators coupled by non-trivial physical mechanisms. This is similar to the study by Pitti et al. (2009), in which simulated biped walkers were coupled to oscillators. There, it was shown that at optimal values of the coupling (where the best walking behaviour is achieved) there is an increase in information transfer from the body to the oscillator. Thus the information transfer is thought to correlate to the successful entrainment of the body and controller dynamics.

Transfer entropy (Schreiber, 2000) is the information gained from conditioning the entropy rate of a time dependent variable on a secondary historical variable as well as its own past. It is a directional measure, and is often interpreted as signifying causal links (Pitti et al., 2009; Lungarella and Sporns, 2006), however when used to analyse finite experimental time series data there is a risk of over- and under-estimating such causal influences (Marschinski and Kantz, 2002; Lizier and Prokopenko, 2010).

The method of calculating transfer entropy presented here uses a novel approach to producing a discrete time series, inspired by the stroboscopic analysis of Schäfer et al. (1998). Other than that, the method follows Marschinski and Kantz (2002) by conditioning on the longest practicable history of the target variable (a requirement that is sometimes neglected).

The effect of oscillator coupling on transfer entropy is investigated in a continuous time system composed of either a single oscillator and passive body model, or a pair of two such systems (see Figure 3.1). The following sections will first introduce the models studied, and later develop the transfer entropy calculation based on time series data from simulations of these models. The relationship between transfer entropy and the dynamical process of synchronisation is discussed. It is argued that a state of synchronisation will not always lead to increased transfer entropy, but during an ongoing process of weak synchronisation transfer entropy will be found, and in such

cases will show causal relationships.

3.2 Model construction

The model developed here is intended to be a minimal first approximation to a physically realizable modular active dynamic walker. That is, is deliberately bears some conceptual similarity to the real robotic control algorithms described later in this thesis, but we focus for now on something much more easy to analyse. The fundamental components of the system are: a chaotic oscillator that can be implemented as an electronic circuit, and a simple mass-spring-damper system analogous to a passive compliant robot body. This structure makes it comparable to the architecture of Pitti et al. (2009), except that the neural controller here is a continuous time analog circuit, rather than a discrete time map, and the physical component does not include a full environment.

3.2.1 Chaotic oscillator

The oscillator design developed by Sprott (2000) and improved on in Kiers et al. (2004) was chosen. It is simple to implement using widely available electronic components, and can easily be tuned to produce chaotic or periodic behaviour. These features mean that although the experiment here was conducted using numerical integration (i.e. in a simulation) at least in principle the same experiment could plausibly be conducted with a physical system. We aim to demonstrate the methods developed here on a system that could potentially be part of a real robotic system (we could, for example, use an analog implementation of this circuit as a component of the gait generator for a real robot, although the robot presented later in this thesis ultimately used a traditional computer to generate oscillations). This section introduces both the ordinary differential equation form of the system, and discusses briefly the potential implementation as an analog circuit coupled to a physical resonant system.

The dynamics of the oscillator in isolation are well documented by Kiers et al. (2004), but it is useful to recap them here. The circuit diagram is given in Figure 3.2. The circuit consists of three op-amp integrators producing anti-derivative signals in a chain, with the output of the final integrator being fed back through a non-linear amplifier and combined with the output of the other integrators with a summing amplifier. The circuit effectively implements the following third-order "jerk" function in which $D(x) = -6\min(x, 0)$ and Q and α are constants.

$$\dddot{x} = -Q\ddot{x} - \dot{x} + D(x) - \alpha \tag{3.1}$$

The constant bias α is provided by the voltage source, and for a range of small positive values

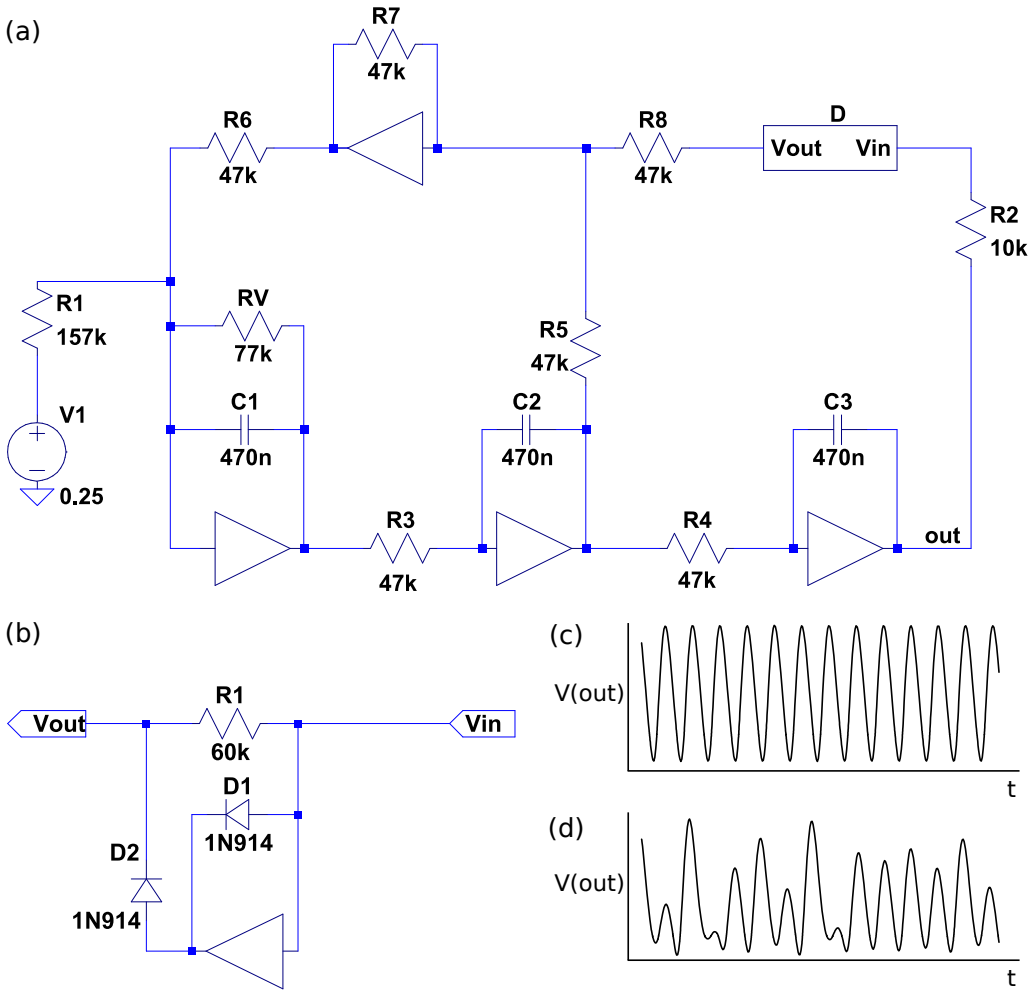


Figure 3.2: Chaotic oscillator based on Kiers et al. (2004). (Resistor values in Ω, capacitor values in F) (a) Positive feedback based circuit design. (b) Non-linear sub-circuit component D. (c) Representative steady state periodic output simulated using LTSpice, with resistor R_v set to $50\text{k}\Omega$ (periodic solution) and (d) $77\text{k}\Omega$ (chaotic solution).

(e.g. $\alpha = 0.1$ works well) it will allow oscillatory behaviour. Tuning R_v in the circuit will vary the Q parameter, which will result in chaotic and periodic solutions at different values, as shown in Figures 3.2c and d, generated by simulating the circuit with LTSpice[1]. The fundamental frequency of the oscillations is related to the time constants of the integrators, that is $\frac{1}{\omega_0} = \tau_0 = RC = 47\text{k}\Omega \times 470\text{nF} \approx 0.022\text{s}$ in the example circuit of Figure 3.2. Equation 3.1 is a non-dimentionalised description of the circuit attained by taking the derivatives with respect to $\omega_0 t$ (so if t is in seconds, equation 3.1 will effectively have a natural frequency of $\omega_0 = 1\text{rad s}^{-1}$). Thus control of the fundamental frequency in the numerical solution of equation 3.1 is achieved, when required, by rescaling the time variable by the desired value of ω_0.

Figure 3.3a shows the effect of Q on the dynamics of the system, by showing the maxima of

[1] http://www.linear.com/designtools/software/#LTSpice

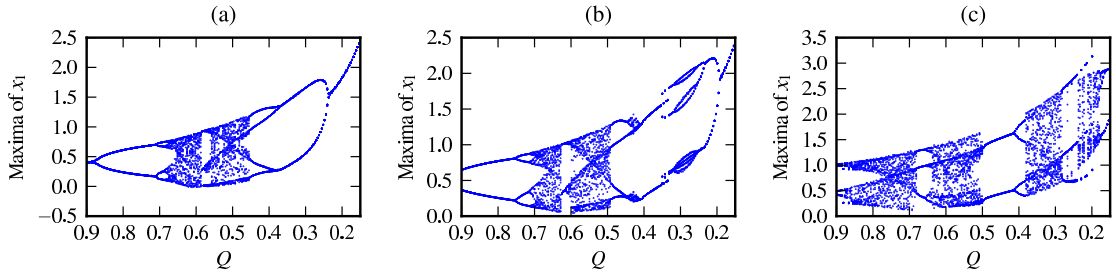


Figure 3.3: Maxima of the oscillator variable x_1 at different values of Q. The fixed parameters are $\alpha = 0.1$, $\zeta = 0.3$ and the coupling is (a) $\gamma = 0$ (no coupling), (b) $\gamma = 0.1$, (c) $\gamma = 0.3$

x_1 over time at different values of Q. The diagram can be obtained by numerically integrating equation 3.1 with a computer library such as LSODE (Hindmarsh, 1983), as was used here, or by using a SPICE simulation of the circuit in Figure 3.2. The system follows a period doubling route to chaos as Q decreases, with a notable periodic island around $Q \approx 0.58$ and returns to periodicity via a further bifurcation near $Q \approx 0.47$.

3.2.2 Coupled mass-spring-damper

The dynamics of an ideal mass-spring-damper (MSD for brevity) can be expressed in terms of the time dependent extension of the spring x using the second order differential equation 3.2 with m being the mass, k the spring constant and c the damping coefficient.

$$\ddot{x} + \frac{c}{m}\dot{x} + \frac{k}{m}x = 0 \tag{3.2}$$

Alternatively define the angular velocity $\omega_0 = \sqrt{\frac{k}{m}}$ and the damping ratio $\zeta = \frac{c}{2\sqrt{mk}}$, take derivatives with respect to $\omega_0 t$ as in equation 3.1 (see above) and rearrange to get:

$$\ddot{x} + 2\zeta\dot{x} + x = 0 \tag{3.3}$$

To couple the two systems together, the oscillator variable is added to the acceleration of the spring system after subtracting 0.5 (to make the influence of the oscillator approximately symmetric around zero, as it normally oscillates between around 0 and 1), and the spring extension is added to the feedback path of the oscillator after multiplying by a coupling parameter γ. Thus the complete system is given by equations 3.4 and 3.5, with x_1 and x_2 being the time varying oscillator variable and spring extension respectively.

$$\dddot{x}_1 = -Q\ddot{x}_1 - \dot{x}_1 + D(x_1 + \gamma x_2) - \alpha \tag{3.4}$$

$$\ddot{x}_2 = -2\zeta\dot{x}_2 - x_2 + (x_1 - 0.5) \tag{3.5}$$

Note that the derivatives in both equations are taken with respect to the same time variable $\omega_0 t$ and thus the oscillator and MSD systems always have identical natural angular velocities. This implies $RC = \sqrt{\frac{m}{k}}$ – that is, the natural frequency of the spring is tuned to the natural frequency of the oscillator (it would be slightly unrealistic to achieve this in a physical implementation of the system, but it will serve as an approximation).

The coupling could be achieved electronically by adding a voltage signal into the input of the non-linear feedback amplifier in the circuit in Figure 3.2 via a series variable resistor such that when the diodes are switched off the op-amp in the non-linear sub-circuit D acts as a summing amplifier and the variable resistor allows control of the coupling strength.

With no coupling ($\gamma = 0$) clearly the oscillator drives the MSD but will not be influenced by it, thus the bifurcations of the oscillator dynamic will remain as in Figure 3.3a. With $\gamma > 0$ the bifurcation structure changes dramatically, as shown in Figures 3.3b and 3.3c.

3.2.3 Synchronisation vs. resonance

Is the change of the dynamics as coupling is increased shown in Figure 3.3 a form of synchronisation? When coupling (here meaning the feedback coefficient γ) is zero an engineer might call the MSD system a passive resonant filter – remember that is it still driven by the electronic oscillator, so the frequency spectrum of the spring extension will appear to be a filtered version of the oscillator output. With feedback however, the oscillator changes its behaviour noticeably as we have seen, so perhaps there is something more than simple resonance happening – a form of synchronisation.

This appears to be the view taken by Pitti et al. (see Figure 3 in Pitti et al., 2009), who suggest that resonance is the forward process (from oscillator to dissipative system) and synchronisation occurs along the feedback path. However this seems to contradict the view of Pikovsky et al. (2001, pp. 14–17) and Ceguerra et al. (2011), who require that synchronisation only applies to synchronous variation of systems that are capable of oscillating independently.

Pikovsky et al. give the example of an ecological system such as hare-lynx populations, where both variables (the populations of the two species) oscillate in a phase locked manner, but the system cannot be decomposed into isolated subsystems. Assuming the lynxes eat only hares, then an isolated lynx population with no access to hares would simply die out, not oscillate at some

natural frequency. Likewise an isolated mass-spring-damper not stimulated by an appropriate oscillator will die down as its energy dissipates.

Of course some mechanical systems can oscillate independently – think of a passive dynamic walker (McGeer, 1990, see also chapter 7), a biped structure that walks down a hill, obtaining its energy purely from gravitational potential. As long as the slope is present, the passive dynamic walker could be considered to have its own natural oscillation. If this were the system being coupled to a neural oscillator, then it would seem that true synchronisation could be discussed. However the current scenario of an isolated mass-spring-damper is unambiguous – there is no decomposition of the combined system that leaves two distinct oscillators and hence no synchronisation as a "complex dynamical process, not a state" (Pikovsky et al., 2001). In the later experiments of Pitti et al. (2009) there are multiple neural oscillators in a single system, so the notion of synchronisation becomes more applicable. This scenario will also be investigated later in this chapter.

3.3 Transfer entropy

This section will develop a measure of transfer entropy that can be meaningfully applied to continuous time oscillators.

Initially, assume we have two discrete time series x and y of finite length. The value of x at time $n \in \{1, 2, \ldots, N\}$ is x_n, discretised such that $x_n \in \mathcal{X}$, where $\mathcal{X}$ is a finite set of symbols. Apply the same notation to y.

The k-history of x at n, i.e. $\{x_n, x_{n-1}, \ldots, x_{n-k+1}\}$ is written $x_n^{(k)}$, and likewise the l-history of y is $y_n^{(l)}$. We treat the time series x and y as samples of equivalent random variables X and Y The transfer entropy from series Y to series X, written $T_{Y\to X}$, is the information gained about X_{n+1} in moving from prior knowledge of $X_n^{(k)}$ alone to also having $Y_n^{(l)}$. This is given by the Kullback-Leibler divergence, or equivalently the conditional mutual information, calculated using the summation in equation 3.6.

$$\begin{aligned} T_{Y\to X} &= I(X_{n+1}; Y_n^{(l)} | X_n^{(k)}) \\ &= D_{KL}(P(x_{n+1}|x_n^{(k)}, y_n^{(l)}) || P(x_{n+1}|x_n^{(k)})) \\ &= \sum_{\mathcal{X}} P(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log \frac{P(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{P(x_{n+1}|x_n^{(k)})} \end{aligned} \tag{3.6}$$

The probabilities are estimated from observations of a single instance over a long time series. This is similar to the methods of Pitti et al. (2009) and Marschinski and Kantz (2002). It is very important therefore that the time series is statistically stationary over the period of interest,

which can be a practical problem with transfer entropy calculations. It is also possible to calculate similar information transfer statistics from ensembles of non-stationary systems by calculating probabilities from the ensemble at each point in time (e.g. Ceguerra et al., 2011; Williams and Beer, 2010a). The time average approach was chosen here because there is at least potential applicability to complex real systems (e.g. a real robot) where experiments cannot be repeated in such a way that the entropy calculation would be possible and valid.

The following sections consider further practical issues regarding the application of this measure to the simulated time series in these experiments. Since these time series are continuous, sensible discretisations must be established. Further, appropriate values of k and l need to be chosen.

3.3.1 Stroboscopic discretisation

The time series are first analysed using a stroboscopic method similar to that of Schäfer et al. (1998). Consider again the continuous time series generated by the coupled oscillator-MSD system in equations 3.4 and 3.5: $x_1(t)$ and $x_2(t)$. The series of maxima of the oscillator voltage are $\hat{x}_1$ and the time of the nth maximum of x_1 is $\hat{t}(n; x_1)$. Figure 3.4 shows the phase of x_2 plotted at each maximum of x_1.

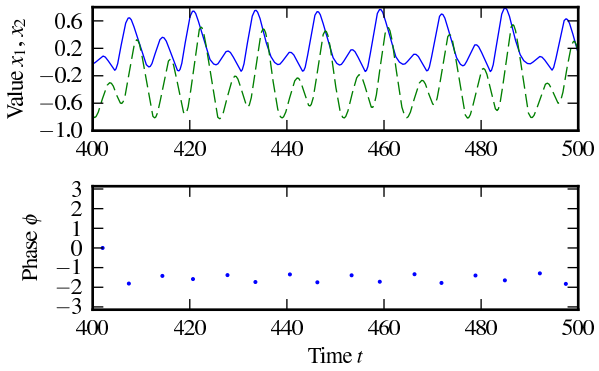


Figure 3.4: Stroboscopic visualization of the spring extension at each maximum of the oscillator in the coupled oscillator-MSD system. Top: The solid blue line is the oscillator voltage x_1, and the dashed green line is the spring extension x_2. Bottom: The points represent the phase of x_2 (in radians) taken at each maximum of x_1. The oscillator is set to a chaotic mode with $Q = 0.67$, the remaining system parameters are $\alpha = 0.1$, $\omega_0 = 1\text{rad s}^{-1}$, $\zeta = 0.3$ and $\gamma = 0.01$.

The phase of the spring extension is calculated here on a "peak-to-peak" basis. That is, the phase of the spring at the nth maximum of the oscillator x_1 is taken to be the (linear) proportion of the time between the last and the next maximum of the spring extension x_2 that has already

elapsed, written $\phi(n; x_2, x_1)$, normalised to the interval $[-\pi, \pi]$. Alternative methods of calculating the phase were considered, such as using the Hilbert transform as per Schäfer et al. (1998). This was found to be problematic due to the chaotic nature of the signals here, hence the simpler peak-to-peak method was chosen, but in other applications the Hilbert transform might well provide useful phase values to use with this transfer entropy method.

For each phase angle of x_2, $\phi(n; x_2, x_1)$ with $n > 1$, the time period of the most recent oscillation of x_1 can also be calculated by:

$$\Delta\hat{t}(n; x_1) = \hat{t}(n; x_1) - \hat{t}(n-1; x_1) \tag{3.7}$$

Thus the two continuous time series x_1 and x_2 can be converted to "stroboscopic" discrete time representation: $\Delta\hat{t}(n; x_1)$ and $\phi(n; x_2, x_1)$, usually defined for all $n \in \{2, 3, \ldots, N\}$ (with the proviso that ϕ is only defined when the nearest maxima of x_2 are known).

It would of course have been possible to simply discretise x_1 and x_2 by choosing arbitrary time intervals. The advantage of the stroboscopic method is that the time intervals are determined naturally by the dynamics of the system. Furthermore, the time series being compared here have a natural interpretation in terms of synchronisation, which is well documented in the literature. In what follows the "stroboscopic" transfer entropy is effectively the influence of the phase difference on the future frequency of oscillation. This will be denoted $ST_{A\to B}$ as shorthand for transfer entropy after the stroboscopic conversion has been applied, i.e. $T_{\phi(\cdot;A,B)\to\Delta\hat{t}(\cdot;B)}$.

3.3.2 Simulation method

The "stroboscopic" time series defined above can easily be obtained from numerical simulation of the oscillator-MSD system. First LSODE was used to obtain a solution to the initial value problem given by equations 3.4 and 3.5 via numerical integration, with the starting values of x_1, x_2 and the necessary derivatives ($\dot{x}_1$ etc) at time $t_0 = 0$ chosen randomly from the range $[0, 1)$. The first part of the time series from t_0 to a chosen cut-off point t_{tr} was discarded to remove the "transient" dynamic of the system. During the following interval, between t_{tr} and the end of the simulation at another chose time t_1, it is assumed that the dynamic reaches an attractor (observation suggests that this is the case).

Inevitably the numerical solution of the equations will give a discrete-time output series, with intervals chosen here to be one twentieth of a (simulated) second between points. Thus oscillations have a period of around 120 simulation intervals (recall that the effective time constant of the system was 2π). The maximum times ($\hat{t}$) were estimated by finding those values in the simulated time series that were preceded and followed by lower values – a crude method but it is effective in

this case.

To estimate the necessary probability distributions, the frequencies of samples in p bins was used, with the bin sizes adjusted such that each has a similar number of data points in it, following Marschinski and Kantz (2002). More advanced methods could be applied but for the current purposes this appeared to be adequate and should produce reliable results.

3.3.3 Causality and transfer entropy

Transfer entropy can sometimes be thought of as "causal information", but care needs to be taken. Transfer entropy is literally, from equation 3.6, an information gain in moving from conditioning the future of X on its own history $X^{(k)}$ alone to conditioning on the joint history of $X^{(k)}$ and $Y^{(l)}$. Suppose that k and l (the history lengths) are both 1, as is sometimes the case in the literature (Pitti et al., 2009; Lungarella and Sporns, 2006). The immediate history $Y^{(1)}$ can contain information about the future states of X without having any real causal influence if it contains information about past states of X that have not been conditioned for in the mutual information calculation – i.e. $X^{(k)}$ is too short a past.

This problem is clearly a possibility in the system under study here. In a single oscillator-MSD system with the feedback coupling γ set to 0, we know from the design of the system that the spring extension has no influence on the oscillator dynamic, but we also know that the MSD system stores mechanical energy and hence contains information about its own past and the past of the oscillator (which stimulated it). Thus the current state of the spring may help to predict the future state of the oscillator when added to just the current state or recent past of the oscillator, but if we control for the entire history of the oscillator, the spring state cannot be useful, as it is itself determined entirely by the history of the oscillator.

Marschinski and Kantz (2002) present a method designed to minimize this overestimation: set l to 1, then increase k from 1 until any causal influence of $X_n^{(k)}$ on the future X_{n+1} is already accounted for before calculating the information gained by including $Y_n^{(1)}$. However, increasing k will rapidly expand $\mathcal{X}$ in equation 3.6, i.e. the support set for which probabilities must be found. In a finite time series this will result in fewer examples of each combination of states from which to calculate the conditional probabilities, and ultimately to sometimes significant overestimation of the transfer entropy. The solution proposed by Marschinski and Kantz is to subtract the transfer entropy obtained when the Y series is randomly permuted in time, so that any true temporal correlations are lost and any calculated transfer entropy must therefore be due to finite sample error. This measure (which they called effective transfer entropy) will be used in all calculations to follow.

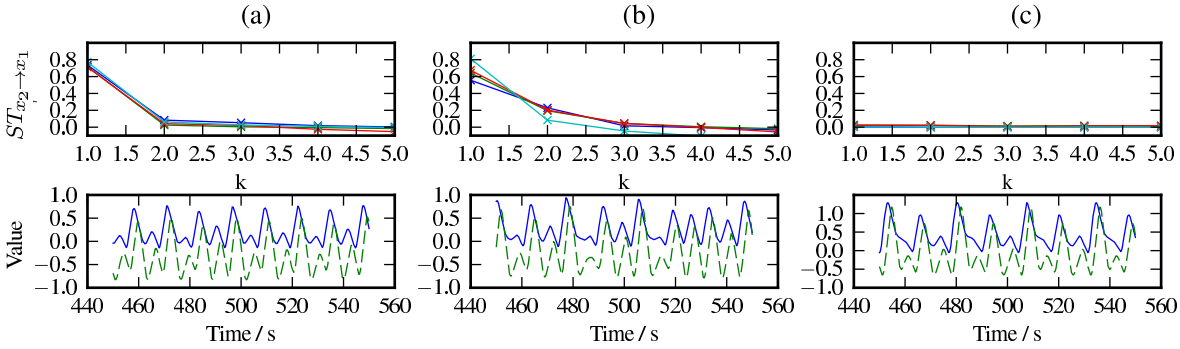


Figure 3.5: Effective stroboscopic transfer entropy for a single coupled oscillator-MSD system. Coupling was (a) $\gamma = 0$ (no coupling) (b) $\gamma = 0.1$ (c) $\gamma = 0.3$. Top row: $ST_{x_2 \to x_1}$ (from spring phase to the oscillator periods), using four partition sizes $p \in \{4, 5, 6, 7\}$. Bottom row: short sections of simulated time series. Other parameters: $\omega_0 = 1\text{rad s}^{-1}$, $\zeta = 0.3$, $Q = 0.67$, $\alpha = 0.1$. Time series analysed between $t_{tr} = 400\text{s}$ and $t_1 = 15000\text{s}$ with measurement interval $\Delta t = \frac{1}{20}\text{s}$.

Figure 3.5a shows how effective transfer entropy overestimates the causal influence of the spring on the oscillator when $k = 1$ even when no feedback coupling is present so the spring cannot possibly have causal influence on the oscillator. Furthermore, it appears that the overestimate at small k is the only source of apparent transfer entropy even when coupling is added: Figure 3.5b, $\gamma = 0.1$, $ST_{x_2 \to x_1}$ rapidly declines when $k \geq 2$. Though the spring does influence the dynamics of the coupled system as we know, this cannot be detected from the time series as the spring can only store information previously generated by the oscillator. When coupling is higher ($\gamma = 0.3$, Figure 3.5c) the system reaches a limit cycle dynamic, and no transfer entropy is detected at any point. The fact that the spring is not an independent oscillator implies both that this is not a true process of synchronisation, and further that no transfer entropy can be measured.

3.3.4 Transfer between two oscillators

The above results show that for transfer entropy to be present with larger values of k (i.e. to genuinely signify causal dependence), there must be at least two systems capable of "producing" information – such that each of the subsystems has a non-zero entropy rate (their future is not trivially predictable from their past state). This can be achieved by duplicating the oscillator-MSD system and coupling via the mechanical component (the MSD). There are therefore two x_1 and x_2 variables (one for each system), the coupling is added by updating equation 3.5 (the dynamics of the MSD) to add the difference between the two springs multiplied by a coupling coefficient γ_c to the acceleration of the local spring. Equation 3.8 therefore gives the acceleration of the local spring $\ddot{x}_2$ given the remote spring extension x_2'.

$$\ddot{x}_2 = -2\zeta\dot{x}_2 - x_2 + (x_1 - 0.5) + \gamma_c(x_2' - x_2) \tag{3.8}$$

This system is now a loose analogue of a pair of mechanically coupled neural-mechanical systems, call them system A and system B. In what follows, system variables and parameters will be superscripted with an A or B to signify (where it is ambiguous) which system they are a part of, e.g. x_1^A and x_1^B are the oscillator values for system A and system B respectively. Usually the parameters will be identical in both systems, in which case a superscript is not used. The introduced coupling γ_c models the mechanical linkage between the two systems, whereas the existing internal coupling γ can be viewed as a kind of proprioceptive feedback to an oscillator from the limb it directly controls. The following experiments will study the effect of changing γ while keeping γ_c constant, i.e. to ask the question: *given a fixed mechanical coupling (body morphology), how do changes in internal coupling affect overall synchronisation, information transfer and causal influences*.

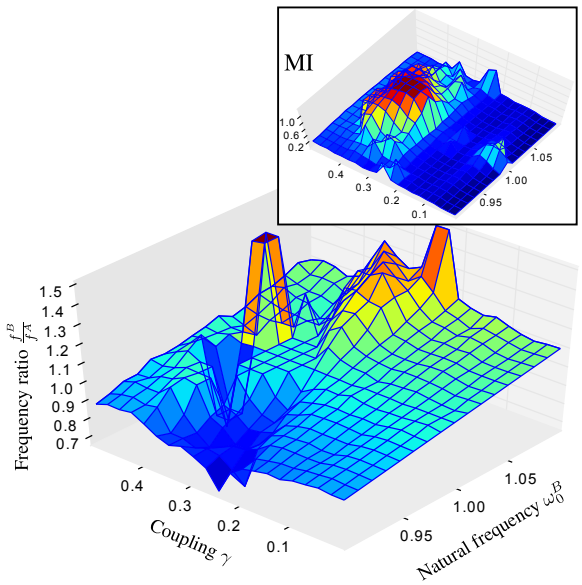


Figure 3.6: Ratio of peak-to-peak oscillation frequency for x_1^A and x_1^B when the natural frequency of the first oscillator is $\omega_0^A = 1\text{rad s}^{-1}$ and the second oscillator is varied near to that, for increasing internal coupling (γ) in both oscillators. The system parameters were $Q = 0.67$, $\alpha = 0.1$, $\zeta = 0.3$, $\gamma_c = 10$, $t_{tr} = 400\text{s}$ and $t_1 = 15000\text{s}$ with measurement interval $\Delta t = \frac{1}{20}\text{s}$. Inset: mutual information between time series x_1^A and x_1^B over the same region in parameter space shows the synchronisation region more clearly.

Figure 3.6 shows how increasing values of internal coupling γ (with the spring coupling fixed at $\gamma_c = 10$) affect the ratio of the mean peak-to-peak frequencies of oscillator B and oscillator A, i.e. $\frac{f^B}{f^A}$ where $f^S = \langle \Delta\hat{t}(\cdot; x_1^S)\rangle^{-1}$. The natural frequency of system A is fixed at $\omega_0^A = 1\text{rad s}^{-1}$

and ω_0^B is varied as shown. Clearly for no coupling the frequencies should vary independently, so the observed frequency difference varies linearly with the natural frequency difference. As coupling is increased, the observed frequencies appear to be pushed further from the natural frequency difference, until at coupling greater than $\gamma \approx 0.3$, a synchronisation region appears around the central part of the plot where the frequencies tend to lock to a 1:1 ratio (except just around $\gamma = 0.45$ where there are two peaks representing areas where the frequency ratio, though still synchronised, is 3:2). As γ approaches 0.5 this region starts to shrink again, suggesting an optimal value of γ (in terms of the likelihood of frequency locking) exists in this region.

The effective transfer entropy when the coupling between the oscillators is $\gamma = 0.35$ is shown in Figure 3.7. The history length was $k = 4$ and $p = 4$ bins were used to discretise each series, the maximum practical values that could be used following the method of Marschinski and Kantz (see above). Mutual information between the instantaneous velocities of the two oscillators (measured by the values of $\dot{x}_1$ produced by the simulation) is used to measure synchronisation, with high values of mutual information implying that the oscillators vary at related speeds and therefore are synchronised. The same binning approach as for transfer entropy is used, but with $p = 5$ bins. The mutual information is also rendered in the inset plot in Figure 3.6, which shows that high mutual information corresponds to the frequency locking region.

The relationship between transfer entropy and synchronisation is complex. There appears to be a main frequency locking region near $\omega_0^B = 1\mathrm{rad\ s}^{-1}$ in Figure 3.7a (where the natural frequencies are most similar) and smaller peaks at larger frequency differences, which are hypothesized to be at points where harmonic resonance along the body allows for greater synchronisation between the oscillators. Note that at the mutual information (synchronisation) peaks, there is usually a trough in the transfer entropy rate, especially in the approximate range $1 < \omega_0^B < 1.05$. Here the synchronisation is strongest, and the transfer entropy is not seen because the two systems are coordinated in a highly synergistic manner, such that the coupling appears to be rigid to an outside observer. Because the systems are not generating entropy independently (i.e. the entropy rate $H(X_{t+1}|X_t^{(k)})$ for either system is 0), no transfer entropy can be measured.

To investigate the notion that the transfer entropy measures directed causal information, the internal coupling γ was set to zero for one of the oscillators (A or B) at a time (with the other retained at $\gamma = 0.35$). Recall that the internal coupling regulates the strength of the signal from the spring extension that is incorporated in the feedback path of the oscillator circuit. Therefore setting the internal coupling to zero for oscillator A will mean that system B cannot have a causal effect on system A, and $ST_{B\to A}$ (the transfer entropy from B to A) should be zero, as shown in Figure 3.7b. Likewise removing the internal coupling from system B results in $ST_{A\to B} = 0$

(Figure 3.7c). In the coupled direction, transfer entropy is generally present. The transfer entropy does not drop to zero in the most synchronised areas of Figures 3.7b and 3.7c as it does in the mutually coupled scenario. This suggests that the synchronisation is weaker and intermittent, allowing the influence of one oscillator on the other to be measured.

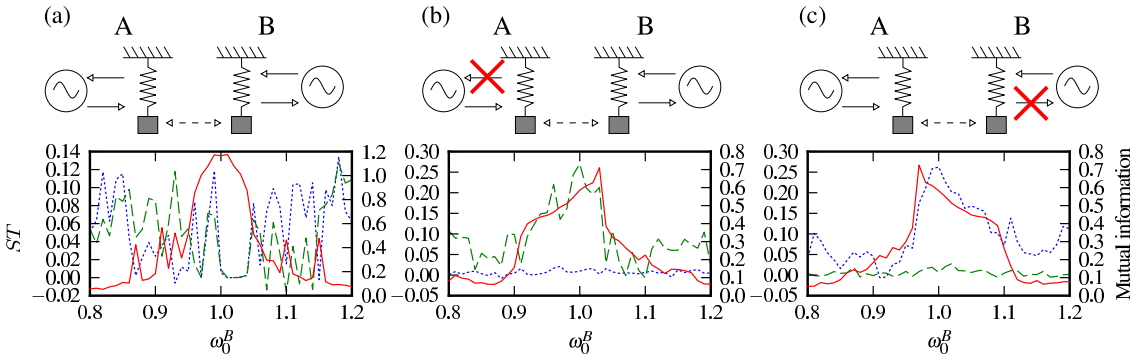


Figure 3.7: Frequency mutual information (red solid line), $ST_{B\to A}$ (blue dotted line) and $ST_{A\to B}$ (green dashed line) for double oscillator system, with oscillator A at $\omega_0^A = 1\text{rad s}^{-1}$ and B at nearby frequencies as shown. Coupling is: (a) mutual, $\gamma = 0.35$ in both systems; (b) no feedback in system A ($\gamma^A = 0$); (c) no feedback in system B ($\gamma^B = 0$). Other parameters as Figure 3.6.

3.4 Conclusions

Transfer entropy from source A to target B is (in a mathematical sense) a measurement of the information about the future of B contained in the state of A, when the past state of B is already known. To infer causality (i.e. "A causes B") we must be sure that the history of B accounts for *all other causal influences on B that may be correlated with A*, particularly the complete history of B (cf. Lizier and Prokopenko, 2010; Ay and Polani, 2008). Properly measured with this taken into account, transfer entropy will be zero if A does not generate information independently of B.

The above has shown that systems that are weakly synchronising are capable of this independent information generation, and thus observational transfer entropy can be measured. Furthermore, transfer entropy is *only* found in the case of weak synchrony, and not for systems that are either not truly synchronising (such as a single mass spring damper coupled to a single driving oscillator), or too rigidly synchronised (as in the case of two very tightly coupled oscillators). Importantly, this means that the observational transfer entropy is not a direct measure of the "strength" of synchrony or causal relationship, because the strongest relationships may show no transfer entropy.

There is a persistent asymmetry in the plots in Figure 3.7 – in the fully coupled scenario, the transfer entropy appears to be generally higher in the direction leaving the oscillator with higher ω_0 (remember that ω_0^A is always 1, thus in the left hand half of Figure 3.7a $\omega_0^A > \omega_0^B$ and

notice that generally $ST_{A\to B} > ST_{B\to A}$). When feedback coupling is removed in one oscillator, synchronisation appears to happen over a larger region when that oscillator has a higher natural frequency, as shown by the asymmetry in the mutual information curves (Figures 3.7b and 3.7c). This relation suggests is may be possible to use the transfer entropy to make useful predictions about the consequences of further interventions, with the important caveat noted above that it cannot be a perfect method of inferring causality.

These fundamental observations about transfer entropy will be of importance in the coming chapters, where we will describe in more detail the connection between information based measures such as transfer entropy and probabilistic models of causation. In particular chapter 4 describes the relationship between transfer entropy and causal graphical models, and chapter 5 discusses further the issue of "non-separability" in the case of strong coupling – the phenomenon seen above where, in the case of a genuine causal influence, transfer entropy may appear to "underestimate" that influence.

Chapter 4

Strength versus inference – a consistent view of information transfer

At the end of the last chapter, we saw that transfer entropy appears to have some genuine correspondence to causal influence, but there was a problem: in the case of strong synchrony, that is, where two time series match each other exactly due to strong mutual influence, transfer entropy would tend towards zero. This chapter aims to give a principled discussion of the relationship between transfer entropy and causal influence. It introduces some of the broader conclusions of this thesis: namely that we can reasonably regard transfer entropy, under specified conditions, as a justified statistic for inferring causal relationships, however, this should not be read as saying that transfer entropy measures the strength of a causal relationship.

A starting point will be to define a precise concept of *information transfer*, bearing some similarity to the definition of *information flow* given by Ay and Polani (2008). The increased precision comes from the use of causal Bayes net theory (see chapter 2) to ground our definitions. Information transfer is a justified statistic for inferring causation, in a sense that we will outline, and can be applied either to time series or to "static" causal relationships between non-temporally specific variables. Transfer entropy, in its typical form, may or may not be strictly a form of information transfer, however provided the history length used in the transfer entropy calculation is long enough, it can be justified as a statistic for causal inference.

We then consider two apparent confounders for the use of information transfer as a tool for causal inference: the first is the failure of the *stability* assumption of the causal graph. The second, specific to time series, is the requirement of *ergodicity*. These issues produce superficially similar problems for the use of transfer entropy, but as we will see they come from substantially different underlying conditions. Both of these points were already briefly discussed in chapter 2, but here we put them specifically in the context of information theoretic statistics. We also see here that with

an appropriate interpretation of information transfer as a tool for statistical inference, we can still obtain an internally consistent view of information transfer in spite of these problems. Indeed, we see that failures of stability and ergodicity primarily confound interpreting information as "causal strength", but this does not mean they prevent its use as a statistics for inferring causation.

In order to fully understand information transfer, we will simultaneously discuss the *information flow* proposed by Ay and Polani (2008). This is a distinct concept which incorporates the concept of *intervention* in the causal graph. Interventions are modelled by removing causal influences in the system that affect the intervened on variables – as a result of considering this, the information flow exposes a genuinely different measure of causal influence. However, we argue that it still should not be seen as a general measure of causal strength – there are cases in which information flow may equally be vulnerable to confounders.

We begin by motivating this inferential perspective by considering a phenomenon that is relevant in the wider context of this thesis, i.e. synchronisation. Synchrony is an example of mutual causal influences, and we will consider how we would infer these causal influences from observation and intervention in section 4.1. Section 4.2 describes this with more mathematical precision, specifically we consider how one infers the presence of "genuine" statistical dependencies in sampled data. In particular, we see how mutual information statistics can be used in this context. To apply this to detecting causal influences, we recall the graphical causal modelling techniques from chapter 2 in section 4.3 – since (by the causal Markov condition) we can in at least some cases derive facts about statistical independence between variables from their causal relationships, we can thus use information as a tool to infer causal influences. Section 4.4 introduces the *information flow* approach of Ay and Polani (2008), and describes how the stability problem may apply to both information transfer and information flow. In section 4.5 we describe how the model can be extended to time series problems, but also see how ergodicity (or lack thereof) may introduce a new type of problem. We also see that even when the assumption of ergodicity is satisfied, transfer entropy does not measure the physical *strength* of causal influence.

4.1 Introduction – sympathetic pendulums

Many of the systems we study in this thesis are instances of synchronising oscillators. An early observation of this phenomenon was made in the 17th century by the Dutch scientist Christiaan Huygens, who was ill in bed when he noticed an "odd kind of sympathy" in two clocks suspended on a common support (Bennett et al., 2002). The pendulums appeared to synchronise such that they were perpetually in exact anti-phase (one reaching its leftwards extremity just as the other reached its rightwards extremity and *vice versa*). This behaviour would continue indefinitely

unless perturbed, but if one of the clocks was knocked out of phase, the system would rapidly resynchronise. The ultimate conclusion was that the two pendulums constantly influenced each other through the common physical connection, a realisation which eventually inspired the diverse modern study of synchronisation in non-linear oscillators (Strogatz, 2004; Pikovsky et al., 2001).

There is an epistemological puzzle here. Putting ourselves in Huygens' shoes, how could we *know* that the pendulums were influencing each other?[1] Possibly because this kind of "sympathy" would seem profoundly unlikely to occur if they were not – we know that it is incredibly hard, if not impossible, to construct two pendulums that swing at *exactly* the same frequency, and equally difficult to start them off at exactly the same time in exactly opposite positions. Even before one is perturbed, we have good reason to believe that there is some separate influence at work. This, I think, bears close relation to Deborah Mayo's conception of reasoning from error (Mayo, 1996): an inference that the clocks influence each is warranted because if that were not the case, the observed behaviour would be extremely unlikely (if not impossible).

But there is a second problem – how do we know that the clocks were actually influencing *each other*, and not simply driven for example by a third clock just out of view? For either explanation, we can easily imagine an experiment where the observable result of two clocks swinging in anti-phase might be seen. However, we have an important piece of background information – assuming we know that a third clock is not present in the experiment, then we can exclude that explanation.

If we cannot exclude such a possibility by other means, then perhaps the simplest solution is to accept that no inference can be made – if we are not able to exclude the possibility of a third clock, then we simply do not exclude it (though we may take steps to exclude it by means of further experiment). We neither rule it out, nor rule it in. We could perhaps state a "probability" of there being a third clock (or any other possible explanation), if we are forced to guess, but nothing otherwise seems to oblige us to do so.

The information theoretic statistics discussed below can be thought of as more mathematically precise instantiations of this type of logic. Provided appropriate background conditions are met, we can describe the probability of a particular statistical result under a certain hypothesis and, in the case that the probability of the result observed is low, exclude a given hypothesis. In the case where necessary background conditions are not met, we are simply required to reserve judgement.

4.2 Severity and discrepancy

This section describes in more detail the process of statistical inference, and in particular discusses the distinction between *severity* and *discrepancy* (to use the terms adopted by Mayo (1996)). The

[1] The question of how Huygens himself worked it out is also of interest, but is a problem for a historical analysis.

following will be based on the overview of error statistics in Mayo and Spanos (2010), wherein the following three concepts are key:

The severity requirement permits inference towards a hypothesis when it has *passed* a severe test, that is, an experiment has been conducted which supports the hypothesis and would be unlikely to produce a similar result if the hypothesis was false.

Accordance refers to how well some data fits a hypothesis. In formal approaches, it is measured by a *test statistic* – something which summarizes the deviation of the *data* or *experimental outcome* from what is expected under a given hypothesis.

Discrepancy between hypotheses – some reasonable measure of how distinct two hypotheses are. When a hypothesis is formally specified by the true value of a parameter, the size of the discrepancy is the difference in the hypothesised value, e.g. if one hypothesis is $H_0 : \beta^* = \beta_0$ and another is $H_1 : \beta^* = \beta_0 + \gamma$ then γ is the discrepancy.

The concept of discrepancy is related to the traditional notion of effect size. Here we argue that while accordances are well defined by statistical requirements, discrepancies are not – they come instead from the model of the underlying system. In non-parametric or model free approaches the discrepancy is usually assumed to be the same thing as the accordance (namely an information measure). This is not in itself unreasonable, but in cases where the discrepancy is more naturally defined by some other quantity (such as some physical causal coupling), the difference between the "natural" discrepancy and the information theoretic discrepancy is a potential source of confusion.

To begin with, we formally define *severity* as follows:

> (Severity, definition quoted verbatim from Mayo and Spanos, 2010) A hypothesis H passes a test T with data x_0 "with severity" if
>
> - (S1) x_0 accords with H,
> - (S2) with very high probability, T would have produced a result that accords less well with H than x_0 does if H was false.

Treating accordance as the value of a test statistic, (S1) is interpreted as stating that the value of this test statistic is close to what it would be under the assumption that H is true. (S2) is evaluated using the sampling distribution of the test statistic *under the alternative hypothesis that H is false.* For example, we could model a system using the simple linear regression:

$$y = \alpha + \beta x + \varepsilon$$

The terms α and β are *parameters* of the model, which we assume have true values α^* and β^*. We are for the sake of argument interested in testing whether β^* is zero or non-zero (and almost entirely uninterested in α^*) – formally we specify two hypothesis $H_0 : \beta^* = 0$ and $H_1 : \beta^* \neq 0$. If we set aside for a moment any possible causal interpretation of the linear regression then the symbol x in the above equation can be treated as known value of a random variable X, and is regarded as an "explanatory" variable for y, an instance of Y. A statistical test is performed after collecting N pairs of x and y values. First we obtain from the data the maximum likelihood estimate of the slope $\hat{\beta}$ and then one of two test statistics, either the Z statistic with a Gaussian sampling distribution:

$$Z = \hat{\beta} / \frac{\sigma}{\sqrt{\sum_n (x_n - \bar{x})^2}} \overset{H_0}{\sim} N(0,1)$$

which is used if we know the exact standard deviation of the error term $SD[\varepsilon] = \sigma$. If we did not know the exact value of σ we use the t statistic, which gives the Student's t distribution with $N - 2$ degress of freedom, N being the number of data points:

$$t = \hat{\beta} / \sqrt{\frac{\sum_n (y_n - \bar{y})^2}{(N-2)\sum_n (x_n - \bar{x})^2}} \overset{H_0}{\sim} t(N-2)$$

Call the calculated value of Z or t from the data Z_0 or t_0. Assuming the calculated statistic is non-zero the data can be said to accord with the hypothesis $H_1 : \beta^* \neq 0$, satisfying (S1) with respect to H_1. The value associated with (S2) is one minus the two-sided "tail area" of the sampling distribution – the probability of all values of Z (t) smaller (less discordant from H_0) than the observed value Z_0 (t_0). In other words, $\mathrm{SEV}(T, Z_0, H_1)$ or $\mathrm{SEV}(T, t_0, H_1)$ – "the severity with which test T passes hypothesis H_1 with data summarised by Z_0 (t_0)" – is $\Pr_{H_0}(|Z| < |Z_0|)$ or $\Pr_{H_0}(|t| < |t_0|)$. (S2) states that the alternative hypothesis H_1 is severely tested to the extent that this probability is close to 1. If the severity associated with (S2) is not sufficiently high, we might say that the observed result is not "statistically significant," meaning that we do not "reject" H_0. This is just one way that the severity criteria (S1-S2) can be used to describe traditional frequentist testing, in this case the Fisherian significance test where the severity (with which H_1 is passed) is one minus the p value.

However we can go further using the concept of discrepancy. Mayo and Spanos (2010) discuss how various objections to frequentist testing can be sensibly countered by combining severity and discrepancy. For example, we can define "substantive significance" as some meaningful discrepancy, say γ. Then we could produce a new pair of complementary hypotheses: $H_0' : |\beta^*| < \gamma$ and $H_1' : |\beta^*| \geq \gamma$. Establishing the severity with which H_1' is tested (which we can do from the same data) tells us whether a substantive discrepancy can be reasonably inferred, avoiding the common

complaint that rejecting the original "point null" H_0 is trivial when the sample size is large since in most cases it is inconceivable that the true β^* is *exactly* zero.

The above example is for a model-based inferences, i.e. we started with a parametric equation and constructed a hypothesis about those parameters. Mayo and Spanos (2010) primarily discuss this type of inference, but non-parametric or model-free tests are also common. Note that in the case of the linear regression, the null hypothesis (zero slope) entailed statistical independence between X and Y. Thus let us construct an equivalent model-free test by beginning with the definition of statistical independence for two variables:

$$H_0 : \forall x, y \in \mathcal{X} \times \mathcal{Y}, \Pr(X = x \wedge Y = y) = \Pr(X = x) \Pr(Y = y)$$

It is well known that the mutual information $I(X;Y)$ is zero when X and Y are statistically independent, and positive otherwise. Thus the mutual information is a candidate statistic for testing this hypothesis – we will consider here how we construct the mutual information derived G test (Sokal and Rohlf, 1991) commonly used to test for statistical dependence between measurements of categorical or discrete-valued variables (i.e. where the sets $\mathcal{X}$ and $\mathcal{Y}$ contain a finite range of distinct values, rather than representing a range of real values as would be typical in the regression model).

First, we will consider a set of artificially constructed "parameters" for the system. Below N data points are collected and n_x is the number of those where $X = x$ (likewise for n_y, and $n_{x \wedge y}$ is the number of data points where $X = x$ and $Y = y$ simultaneously). We expect naive estimates of the probability of each X and Y (and combination X, Y) value to converge to the true probabilities by the law of large numbers ($\doteq$ means equal to in the limit of large N):

$$\begin{aligned}
\hat{\theta}_x &= \frac{n_x}{N} \doteq \Pr(X = x) = \theta^*_x \\
\hat{\theta}_y &= \frac{n_y}{N} \doteq \Pr(Y = y) = \theta^*_y \\
\hat{\mu}_{xy} &= \frac{n_{x \wedge y}}{N} \doteq \Pr(X = x \wedge Y = y) = \mu^*_{xy}
\end{aligned}$$

Then rewrite the hypothesis:

$$H_0 : \forall x, y \in \mathcal{X} \times \mathcal{Y}, \mu^*_{xy} = \theta^*_x \theta^*_y$$

We now want a test statistic based on the maximum likelihood estimates $\hat{\mu}_{xy}$, which (as above) were obtained from the finite sample of N data points. These which would accord best with the null hypothesis when $\hat{\mu}_{xy} = \hat{\theta}_x \hat{\theta}_y$ (since we do not know the true values θ^* we use the estimates $\hat{\theta}$ as *ancillary* statistics, Fisher, 1955). Recall that the mutual information under H_0 is zero:

$$I_{H_0}(X;Y) = \sum_{x,y} \mu^*_{xy} \log \frac{\mu^*_{xy}}{\theta^*_x \theta^*_y} = \sum_{x,y} \mu^*_{xy} \log \frac{\theta^*_x \theta^*_y}{\theta^*_x \theta^*_y} = 0 \tag{4.1}$$

The estimated mutual information $\hat{I}(X;Y)$ is calculated analogously to the true mutual information:

$$\hat{I}(X;Y) = \sum_{x,y} \hat{\mu}_{xy} \log \frac{\hat{\mu}_{xy}}{\hat{\theta}_x \hat{\theta}_y} \tag{4.2}$$

usually this is then normalised by multiplying by $2N$ to get a statistic we will call G, which is known to be distributed as χ^2 under the null hypothesis (that X and Y are statistically independent) (Kullback, 1959):

$$G = 2N\hat{I}(X;Y) = 2N \sum_{x,y} \hat{\mu}_{xy} \log \frac{\hat{\mu}_{xy}}{\hat{\theta}_x \hat{\theta}_y} \overset{H_0}{\sim} \chi^2(k)$$

The degrees of freedom k is $(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$. The expectation of G under H_0 is as a consequence $2Nk$. Under some alternative hypothesis H_1 which entails a true mutual information of $I_{H_1}(X;Y)$, G would be distributed as non-central χ^2 with a centrality parameter of $2NI_{H_1}$ and hence have an expectation value of $2N(k+I_{H_1})$. Note that mutual information cannot be negative, and assuming that H_1 predicts statistical dependence I_{H_1} will be strictly positive. This makes G suitable as a measure of accordance and as a tool for inference via (S1-S2) – lower values accord better with the hypothesis of statistical independence (H_0), and the severity (when we pass the alternative hypothesis of statistical dependence, say) can be calculated given the known sampling distribution under H_0.

What about discrepancy? Before, we defined discrepancy by changing the parameter value claimed in the hypothesis. Now there are multiple parameters in the hypothesis $(\mu^*_{xy}, \theta^*_x, \theta^*_y)$, we presumably need to account for all of them and summarise the discrepancy somehow. There are, fairly obviously, numerous ways of doing this, and a reason for picking one over the other may depend on the task at hand. Suppose that after collecting data we propose some alternative hypothesis H_1 which specifies an alternative set of values for the parameters: $H_1 : \mu^*_{xy} = \mu'_{xy}$. If this hypothesis is discrepant from H_0 in terms of statistical independence, but not anything else, then we should choose the implied marginal distributions to be consistent with the known ancillaries, i.e. choose μ'_{xy} such that $\theta'_x = \sum_y \mu'_{xy} = \hat{\theta}_x$. A possible choice for measuring discrepancy is the Kullback-Leibler divergence between the distribution functions f implied by H_0 and H_1:

$$D_{KL}(f_{H_1}||f_{H_0}) = \sum_{x,y} f_{H_1}(x,y) \log \frac{f_{H_1}(x,y)}{f_{H_0}(x,y)} = \sum_{x,y} \mu'_{xy} \log \frac{\mu'_{xy}}{\hat{\theta}_x \hat{\theta}_y} = \sum_{x,y} \mu'_{xy} \log \frac{\mu'_{xy}}{\theta'_x \theta'_y}$$

But this is just the "true" mutual information implied by the discrepant hypothesis H_1 – $I_{H_1}(X;Y)$. This is not a completely unreasonable choice of discrepancy measure, but it is not obviously or necessarily the correct one either. In the linear regression case, the measure of discrepancy corresponds to some model parameter, where by "model" we mean some physical or otherwise intuitively interpretable mathematical representation of the system under study. In the non-parametric case, the discrepancy has been defined by artificially constructing parameters, which I will call the *statistical parameters*, as opposed to the *model parameters*, until we end up with the discrepancy in the same mathematical form as the accordance. However, there may be some alternative form of the model (e.g. non-linear structural equations), where model parameters *do exist* and could be used to create a natural measure of discrepancy, but these *model parameters* are not the same as the *statistical parameters* used the information theoretic measure. The result is that a discrepancy defined in terms of model parameters may be completely different to a discrepancy defined in terms of statistical parameters.

A key point here is that the discrepancy should be a measure that is relevant and meaningful from the point of view of the scientific theory which the inference is meant to address. As noted above, it could be used to describe *substantive* significance – where substantive means a discrepancy above some threshold which we regard as meaningful for reasons which are generally specific to the problem at hand and not defined by statistical convenience. This constraint does not apply to the accordance or test statistic, which is generally chosen precisely for its statistical properties.

4.2.1 Neyman-Pearson tests

The Neyman and Pearson (1933) (NP) formulation of significance tests can also be compared to the severity conditions. Remember that the severity is evaluated *post-data* – that is, it tells us specifically how strongly the observed data indicate an error in a hypothesis. On the other hand, the NP formalism is based on the properties of the tests themselves (i.e. without the data being known). This will be convenient for our purposes as we will see later.

In the NP framework, a test is first formulated by specifying two hypotheses, H_0 and H_1. A test statistic is found which measures the accordance with H_0 in the same way as before. A threshold value C is then chosen, and if the test statistic is found to be above this the test result will be said to accord with H_1, otherwise it accords with H_0 – hence we have a "pre-planned" decision procedure about which hypothesis to pass (note that neither the Fisherian test, nor the severity requirement itself, necessarily require this). We then define two probabilities based on the sampling distributions: the Type I error rate α is the probability that a test statistic greater than

the threshold will be observed when H_0 is true: $\Pr_{H_0}(t \geq C)$[2] (where t is the test statistic); the Type II error rate β is the probability of a result that does not reach the threshold when H_1 is true: $\Pr_{H_1}(t < C)$.

In the case where the observed test statistic exceeds the threshold $t_0 \geq C$, we say the the result accords with H_1 (S1), and the severity (S2) is bounded below by $1 - \alpha$ as follows:

$$\begin{aligned} SEV(T, t_0, H_1) &= \Pr_{H_0}(t < t_0) \\ &= 1 - \Pr_{H_0}(t \geq t_0) \\ &\geq 1 - \Pr_{H_0}(t \geq C) \\ &= 1 - \alpha \end{aligned}$$

That is, using the fact that we know $t_0 \geq C$ and hence $\Pr_{H_0}(t \geq t_0) \leq \Pr_{H_0}(t \geq C)$ we find that the severity is bounded below by $1 - \alpha$. If we had $t_0 < C$ we would pass H_0 and find by similar reasoning that the severity is at least $1 - \beta$ (the value $1 - \beta$ is known as the statistical *power* of the test).

Supposing we have a mutual information test where H_0 specifies statistical independence as above, α can then be calculated for a given C using the χ^2 sampling distribution. Assuming an alternative hypothesis H_1 is specified which predicts mutual a mutual information value of $I_{H_1}(X;Y)$ then the sampling distribution of the G statistic under H_1 is a non-central χ^2 distribution with centrality parameter $2NI_{H_1}(X;Y)$, hence β can also be calculated.

The key difference between the NP framework and the severity approach is that we can calculate α and β pre-data, if we assume that a decision cut-off C is fixed, whereas severity is only defined post-data and allows the particular result (rather than its relation to a cut-off) to be considered. Following Mayo and Spanos (2010) there are good reasons to generally prefer the severity approach, but the NP framework can be used to describe tests themselves (rather than experimental results) which will be useful below when we consider hypothetical scenarios where no data have actually been collected.

It also helps to give us an intuitive understanding of the "true" mutual information predicted by a hypothesis I_{H_1}. The power $1 - \beta$ increases as I_{H_1} increases (figure 4.1), thus, if H_1 predicts a higher information than H_0 then the two hypotheses are *easier to distinguish* by experiment.

[2] $\Pr_{H_0}(t \geq C)$ means the probability of t taking a value greater than C according to the model implied by H_0. This might also be written $\Pr(t \geq C; H_0)$, but note that we avoid the "conditional probability" formulation $\Pr(t \geq C|H_0)$ since that implies the existence of a "probability" for the hypothesis itself (since $\Pr(t \geq C|H_0) = \Pr(t \geq C \wedge H_0)/\Pr(H_0)$ by definition). Frequentist methods eschew assignment of probabilities to hypotheses.

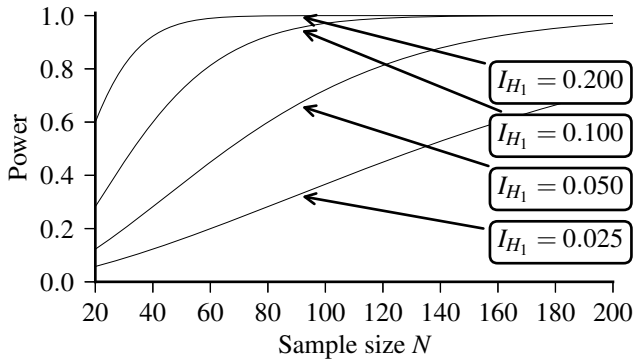


Figure 4.1: The power $1 - \beta$ increases with sample size N and mutual information I_{H_1}.

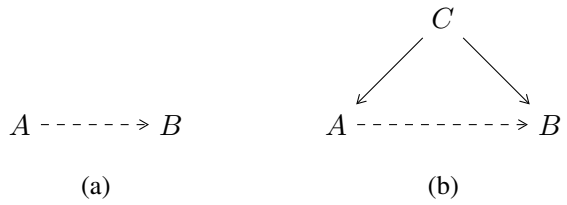
Furthermore, this means that if we do not reject H_0, we can say that we have good evidence for H_0 only in those cases where H_1 predicts a high mutual information (and thus the severity is high).

The above discussion was limited to statistical considerations without reference to the possible causal interpretation of tests. In the following sections we will extend the argument to see how mutual information can be justified as measuring the accordance with a hypothesis of no causal influence between two variables.

4.3 Inference of causal influence

This section will adopt a directed acyclic graph (DAG) formalism for causal models. That is, we regard the statement A causally influences B as equivalent to the presence of a directed path from A to B in a graph (Spirtes et al., 2001; Pearl, 2009). This section regards a *causal hypothesis* as corresponding to a possible *direct* causal influence between two variables, that is, a single arrow from one variable to another. Formally, define a *query graph* for hypothesised influence H as a DAG $G_H = (\mathbf{V}, \mathbf{E}, H)$. The graph *nodes* $\mathbf{V}$ are identified with a set of random variables, such that the variables are assigned same symbol set $\mathbf{V}$ as the nodes, since there is no need generally to maintain a distinction. There is a set of "known" edges $\mathbf{E}$ and a single hypothetical edge H. These edges are ordered pairs of nodes / random variables, for example, Figure 4.2 shows two query graphs: (a) $(\{A, B\}, \emptyset, (A, B))$ and (b) $(\{A, B, C\}, \{(C, A), (C, B)\}, (A, B))$.

The *complete graph* is the graph G_C which represents the "true" causal relationships in the system (i.e. the "real" system graph which we do not yet know but wish to infer). There are two possibilities – either the influence H is present in G_C or it is not. Write this as two competing hypotheses:



(a) (b)

Figure 4.2: Two simple causal networks. Anything not represented in the graph is assumed to have a negligible causal relationship with anything inside the graph. The dashed arrow from A to B represents the possibility of an effect of A on B. In (a) that is the only possible causal link, in (b) we consider C as a common cause of both A and B.

$$H_0 : G_C = (\mathbf{V}, \mathbf{E}) \tag{4.3}$$

$$H_1 : G_C = (\mathbf{V}, \mathbf{E} \cup H) \tag{4.4}$$

We can build probabilistic models compatible with the two possible G_C graphs implied by the simple query graph in Figure 4.2a by assuming two independent random "noise" variables ε_a and ε_b and assigning the realisations of A and B from functions of the two noise variables:[3]

$$\begin{array}{lll} H_0 : & a \leftarrow f_a(\varepsilon_a) & b \leftarrow f_b(\varepsilon_b) \\ H_1 : & a \leftarrow f_a(\varepsilon_a) & b \leftarrow f_b(a, \varepsilon_b) \end{array}$$

The two models above are *compatible* with the two possible complete graphs G_C implied by the two hypotheses because the assignment to a given variable, say x, depends only the parents of X in the graph and the independent noise source ε_x. If we assume the noise variables ε are mutual independent (this is the Markovian assumption used by Pearl (2009) and already discussed in chapter 2) it is clear that H_0 entails statistical independence between A and B and thus $I_{H_0}(A; B) = 0$, therefore we should be able to use the G test outlined above for testing these two scenarios against each other. This is the simplest type of causal inference we might want to make, where we have assumed that there are no possible confounding factors that would lead to statistical dependence between the two variables if one is not causing the other. This may be because we have randomised variable A (to ensure that ε_a is genuinely independent of B), or because of other background assumptions. That is, this may be a model of an *interventional* test or and *observational* one, it is not important provided the background assumptions are correct.

[3]The assignment operator $\leftarrow$ is used here to clarify the causal directionality of the model, but the approach is fundamentally the same as the structural equation model formalism used by Pearl (2009), where $=$ would be used in place of $\leftarrow$.

4.3.1 Conditioning out common causes

The key utility of the graphical formalism for the methods being proposed in this paper is to generalise tests of causal hypotheses to more complex situations, for example where there are common causes that we are already aware of. Essentially, we would like to ensure that the hypothesis of no causal influence (H_0) entails a true mutual information statistic with value zero. The happens for the conditional mutual information $I(A; B|\mathbf{C})$ provided that the set of "conditioning" variables $\mathbf{C}$ meets the *d-separation* criteria for A and B in the causal graph predicted by H_0.

In a nutshell, we make A and B statistically independent by conditioning on their common causes, without also conditioning on their common effects. The d-separation criteria (Pearl, 2009) encapsulates this requirement formally: say that A and B are "blocked" or d-separated by $\mathbf{C}$ if, for all paths between A and B:

1. the path contains a chain $X \to Y \to Z$ or a fork $X \leftarrow Y \to Z$ with $Y \in \mathbf{C}$, **or**
2. the path contains a collider $X \to Y \leftarrow Z$ with neither Y nor any descendant of Y in $\mathbf{C}$.

Write "A is d-separated from B in the graph G by the node set $\mathbf{C}$" as $(A \perp\!\!\!\perp B|\mathbf{C})_G$. Given a probabilistic model of the form introduced in the previous section which is compatible with the graph and has an associated joint distribution over all variables $P(\mathbf{v})$ it can be shown that A and B will be conditionally independent given $\mathbf{C}$ in the probabilistic model if the d-separation criteria are met for the same variables:

$$(A \perp\!\!\!\perp B|\mathbf{C})_G \implies (A \perp\!\!\!\perp B|\mathbf{C})_{P(\mathbf{v})} \tag{4.5}$$

Where $(A \perp\!\!\!\perp B|\mathbf{C})_{P(\mathbf{v})}$ signifies conditional statistical independence in the probability model. This result is the causal Markov condition (CMC) already seen in chapter 2 and is necessarily true if the noise terms are mutually independent Pearl (2009). Recall also from chapter 2 that there is a converse implication

$$(A \perp\!\!\!\perp B|\mathbf{C})_{P(\mathbf{v})} \implies (A \perp\!\!\!\perp B|\mathbf{C})_G \tag{4.6}$$

known as *stability*. However, as was discussed, this is a distinct assumption to CMC and does not necessarily hold everywhere that CMC is true (even though it is not uncommon for it to be assumed to be true in causal inference procedures).

Finally, conditional statistical independence entails a conditional mutual information of zero:

$$(A \perp\!\!\!\perp B|\mathbf{C})_{P(\mathbf{v})} \iff P(a|b, \mathbf{c}) = P(a|\mathbf{c}) \iff I(A; B|\mathbf{C}) = 0 \tag{4.7}$$

Combining equations 4.5 and 4.7, we can say that "d-separation implies zero conditional mutual information". This is the basis of of mutual information tests for causal influence. The manner of this is perhaps already sufficiently implied by the above, but I will try to be specific.

Given a query graph $G_H = (\mathbf{V}, \mathbf{E}, H)$ take the hypothesis H_0 as above to state that the true system is compatible with the complete graph $G_C = (\mathbf{V}, \mathbf{E})$ where the putative edge H is not present. Without loss of generality the edge is $H = (A, B)$, and if we have measurements of a set of variables $\mathbf{C}$ that d-separates A and B then we know that $I_{H_0}(A; B|\mathbf{C})$ is zero by equation 4.5. The contrary hypothesis H_1 that $G_C = (\mathbf{V}, \mathbf{E} \cup H)$ must permit (though does not require without the stability/faithfulness assumption) a non-zero true mutual information $I_{H_1}(A; B|\mathbf{C}) \geq 0$, since the edge H introduces a path from A to B that is cannot be blocked by $\mathbf{C}$.

So a greater value of an estimated information $\hat{I}(A; B|\mathbf{C})$ is more discordant with H_0 which we know entails a mutual information of zero. Thus low values would lead to inference of H_0 and high values to inference of H_1 according to (S1). But of course to establish the severity with which the inference is obtained, we need to consider (S2) – the probability of a result as or more accordant with the inferred hypothesis than the given one, assuming the other hypothesis is in fact true. The complications arising from this will be more clear when we look at some examples. First we will consider what (if any) difference an intervention makes.

4.4 Information transfer vs. flow

Again consider the query graph in figure 4.2b. Now suppose we randomised A using another random variable U which we ensure is independent of all the existing ones. That is, for each draw u of U we $do(A = u)$ – the $do(\cdot)$ notation is introduced by Pearl (2009) to describe this kind of intervention. Following Pearl (2009), we consider how this intervention changes the causal model: first, in the probabilistic model we replace the assignment to a from a function of c

$$a \leftarrow f_a(c, \varepsilon_a)$$

with an assignment to whatever value is chosen for u:

$$a \leftarrow u$$

In the graphical model, we can treat this intervention as removing all the edges terminating at A, and adding a new edge (U, A). In the post-intervention model, the empty set now d-separates A from B as a consequence of this graph "surgery". Thus the mutual information $\hat{I}(A; B)$ *obtained from the interventional experiment* is also a valid test statistic for the same hypothesis that A does not causally influence B.

Choose the marginal distribution of the "control" variable U to be identical to the marginal distribution of A, i.e. assuming the supports of U and A are the same, let $P(u) = P(a)$ when $u = a$. The mutual information $I(A; B)$ in this carefully configured interventional model is the information *flow* as introduced by Ay and Polani (2008). That is, define the information flow $IF(A \to B)$ as $I(A; B)$ when A is randomised but has the same marginal distribution as it does when it is not intervened on, this can be written in terms of causal effects:

$$IF(A \to B) = \sum_{a,b} P(a)P(b|\widetilde{a}) \log \frac{P(b|\widetilde{a})}{\sum_{a'} P(a')P(b|\widetilde{a}')} \tag{4.8}$$

Here, we use $P(b|\widetilde{a})$ as shorthand for $P(b|do(A = a))$. In the general case, we can also intervene on a set of additional variables $\mathbf{C}$ in the same way and get a conditional information flow:

$$IF(A \to B|\widetilde{\mathbf{C}}) = \sum_{a,b,\mathbf{c}} P(\mathbf{c})P(a|\widetilde{\mathbf{c}})P(b|\widetilde{a}, \widetilde{\mathbf{c}}) \log \frac{P(b|\widetilde{a}, \widetilde{\mathbf{c}})}{\sum_{a'} P(a'|\widetilde{\mathbf{c}})P(b|\widetilde{a}', \widetilde{\mathbf{c}})} \tag{4.9}$$

The conditional information flow $IF(A \to B|\widetilde{\mathbf{C}})$ is the mutual information $I(A; B|\mathbf{C})$ in the experiment where A and $\mathbf{C}$ are randomised.

The information flow works as a test statistic in a similar fashion to the conditional mutual information. In a query graph where $H = (A, B)$ is the only path from A to B (figure 4.2a), if the hypothesised link is not real, then clearly $IF(A \to B)$ must be zero under H_0 since, having removed all the edges terminating at A, the empty set d-separates A from B in the post-intervention model – in this case, the post-intervention model is in fact no different to the pre-intervention model, and $IF(A \to B) = I(A; B)$. If there is however another path from A to B (not just the direct one hypothesised, but another such as $A \to C \to B$) then the information flow $IF(A \to B|\widetilde{C})$ both randomises and conditions on the intermediate variable, guaranteeing it to be zero when $H = (A, B)$ is not present in the true system.

We have now two test statistics, the first one (introduced in the previous section) is calculated as the standard conditional mutual information, e.g. $I(A; B|\mathbf{C})$, but has additional requirements with respect to the query graph, namely that the hypothesised link H connects A and B and that if H is not present then $\mathbf{C}$ must d-separate A and B. When this condition is met, we will call it information *transfer*. Information flow is also an information statistic, but implies a set of changes to the causal graph that guarantees that it is a suitable statistic for inferring a causal influence in the sense described above. However, these changes do *not* guarantee that information flow is a more powerful test in the Neyman-Pearson sense than the equivalent information transfer statistic, as we see in the following example.

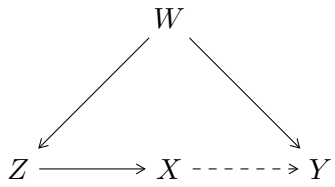


Figure 4.3: Example causal graph.

4.4.1 Example

Take the causal model in Figure 4.3, which shows a query graph where H_0 would specify that the link (X, Y) is not part of the complete graph G_C. Let us consider the case where (X, Y) is in fact present in G_C, i.e. the best inference would be to reject H_0.

We can do this by constructing a system modelled by Bernoulli processes denoted $B(q)$ where the output is 1 with probability q (and 0 with probability $1 - q$). Write the system as a set of assignments:

$$w \leftarrow B(qw) \qquad z \leftarrow B(qz_w)$$
$$x \leftarrow B(qx_z) \qquad y \leftarrow B(qy_{x,w})$$

The parameters give the conditional probabilities of each variable taking a 1 given the value of the parents of that variable in the model. Thus qw is simply a scalar giving the marginal probability $\Pr(W = 1)$, qz and qx are vectors and qy is a matrix with the elements zero-indexed, such that for example $\Pr(Z = 1|W = 0) = qz_0$ and $\Pr(Y = 0|X = 1, W = 0) = 1 - qy_{1,0}$. We set the parameters to the following values:

$$qw = 0.5 \qquad qz = \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix}$$
$$qx = \begin{pmatrix} 0.5 \\ 0.8 \end{pmatrix} \qquad qy = \begin{pmatrix} 0.3 & 0.8 \\ 0.7 & 0.4 \end{pmatrix}$$

The parameters have been arranged so that all the causal links in the graph (including the hypothetical one) can be reasonably described as "real" links – i.e. the value taken by a parent variable clearly affects the conditional distribution of its children, and the correct graph to infer would be the full one with H as a genuine causal connection. However, the parameters have also been chosen such that the probabilistic model violates the stability/faithfulness assumption with respect to this true graph, i.e. when H is present X is not d-separated from Y by the empty set

Table 4.1: Information flow and mutual information calculated for the example system.

Information measure	Exact value (nats)	Power at $\alpha = 0.05$ (n=50 observations)	Degrees of freedom
$IF(X \to Y)$	0.000	0.05	1
$IF(X \to Y \vert \widetilde{Z})$	0.000	0.05	2
$IF(X \to Y \vert \widetilde{W})$	0.076	0.70	2
$IF(X \to Y \vert \widetilde{Z}, \widetilde{W})$	0.069	0.53	4
$I(X;Y \vert Z)$	0.044	0.45	2
$I(X;Y \vert W)$	0.720	0.67	2
$I(X;Y \vert Z,W)$	0.700	0.53	4

(since H constitutes a connection from X to Y) but the two variables are nonetheless statistically independent. This arises because W reverses the effect that X has on Y.

Table 4.3 shows a set of information measures which are all justified as test statistics in the sense I have discussed. The values given are the exact values expected under the hypothesis that the above given parameters are the true parameters for the system, call it H_1. The other hypothesis under consideration, H_0, is that the edge $X \to Y$ is not present in the true causal graph, and hence all the information measures would be expected to be zero under H_0.

Two of the information flow measures $IF(X \to Y)$ and $IF(X \to Y|\widetilde{Z})$ are expected to be zero. This is a problem for using these measures for inference: if we use the NP testing approach, tests based on these measures have minimal power, which is to say they will usually indicate H_0 even when H_1 is true. From the severity perspective, whatever value for these statistics that is obtained by measurement will necessarily accord equally well with the two hypotheses, making it impossible to claim that one hypothesis can be inferred in preference to the other.

By contrast the non-interventional information transfer using Z as a variable that d-separates X from Y gives a positive (albeit small) value for $I(X;Y|Z)$ as desired. The point to stress here is that the information transfer $I(X;Y|Z)$, though not particularly ideal, is a better tool for inference than the information flow defined on the same variables $IF(X \to Y|\widetilde{Z})$, counter to the intuition that the interventional nature of the information flow will makes it necessarily more applicable to causal inference. This is because *whichever causal hypothesis is true*, the information flow will usually give a small value. The information transfer using W – $I(X;Y|W)$ – has a higher exact value than $(X;Y|Z)$, because conditioning on W means the strong dependence of Y on X when W is known can be easily detected. Because Z is influenced by W under the observational regime,

$I(X;Y|Z)$ can likewise detect the same dependence albeit less effectively, but $IF(X \to Y|\widetilde{Z})$ removes the dependence of Z on W, thus making it unable to detect the influence of X on Y.

Of course, since Z also influences (and thus contains information about) X we also have that $I(X;Y|Z,W) < I(X;Y|W)$, thus the information transfer statistic does not necessarily become more powerful as more variables are added, and neither does the information flow since in this case $IF(X \to Y|\widetilde{Z},\widetilde{W}) < IF(X \to Y|\widetilde{W})$.

This example is somewhat artificial and some might suggest even contrived. The point however is merely to demonstrate that the usefulness (for inference) of the information measures depends largely on which variables are conditioned out and not necessarily on whether an intervention has been performed, as can be seen from the range of values in Table 4.1.

This is an example of a violation of the stability assumption, albeit a more subtle one than the birth control-thrombosis example seen in chapter 2. However, the same arguments regarding stability apply. Some would argue that such a parametrisation is in some sense unlikely – recall from chapter 2 the argument that precise choices of the parameters that violate stability (such as the ones chosen above) have Lebesgue measure zero over the space of all parameters. Not only is this argument questionable in its own right, small changes could still be made to the parameters without qualitatively changing the result. For example, with

$$qy = \begin{pmatrix} 0.31 & 0.8 \\ 0.69 & 0.4 \end{pmatrix}$$

the system does, strictly speaking, meet the stability assumption – X is no longer statistically independent of Y. However, this change has a negligible impact on the power of the different statistics – the information flow $IF(X \to Y|\widetilde{Z})$ will still provide a less useful statistic than the information transfer $I(X;Y|Z)$. Thus it is quite conceivable that there is a range of values for the parameters for which intervention would not improve the inferential power of the results.

4.5 Causal influence in time series

The previous section showed that two forms of estimated conditional mutual information, "information transfer" and "information flow" are useful and general statistics for evaluating causal hypotheses provided certain causal assumptions are met. It has also shown that neither can be considered generally more useful for the purposes of inference. This section will consider these measures specifically in terms of time series.

Time series are worth discussing separately because although it is generally quite easy (as we will see) to create "dynamic" causal DAGs that model time series scenarios, they can introduce a

distinct set of assumptions that are required.

We begin by defining a temporal causal model for an m-dimensional system as follows:

- At each time $t \in \mathbf{Z}$ the m-variable random vector $\mathbf{X}_t$ represents the system state. Treating t as an integer gives discrete time steps, though these may in practise be discrete samples of a continuous time process. The state at location i and time t is represented by a random variable $X_{i,t}$. The complete set of variables is a vector $\mathbf{V} = \langle \dots \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2 \dots \rangle$ and there is a set of causal influences $\mathbf{E}$.

- The system must be **temporally Markovian** (distinct from the CMC Markov requirement above). This means the state vector at time t is probabilistically determined by the state at time $t-1$ and thus independent of all previous time steps, i.e. for all $n \in \mathbb{N}$, we have $(\mathbf{X}_t \perp\!\!\!\perp \mathbf{X}_{t-1-n} | \mathbf{X}_{t-1})_G$ which implies $(\mathbf{X}_t \perp\!\!\!\perp \mathbf{X}_{t-1-n} | \mathbf{X}_{t-1})_{P(\mathbf{v})}$. Further, there are no "instantaneous" interactions between variables at the same time t.

- The system must be **causally time invariant** in the sense that there is a causal influence $X_{i,t-1} \to X_{j,t}$ at a given value of t if and only if the exact same causal influence exists for all $t \in \mathbb{Z}$. This means that there is a corresponding edge in the causal graph at all times, and the edge represents the same data generating relationship.

- We apply a **topological constraint** to the causal model by defining a distance between any two random variables across one time step $||X_{i,t-1} - X_{j,t}||$ and specifying a bound on that distance. Only variables which satisfy the constraint are considered as having a possible causal influence edge connecting them in the model.

A typical topological constraint is the "light-cone" found in 1D cellular automata or random fields (cf. Shalizi, 2003; Lizier and Prokopenko, 2010), i.e.

$$||X_{i,t-1} - X_{j,t}|| = |i - j| \leq 1$$

The constraint is illustrated by Figure 4.4 – the light-cone represents a spatio-temporal constraint on the possible causal influences, corresponding to some assumed maximum speed for transmission of causal effects.

Of the edges allowed by the topological constraint, we wish to know which ones really exist. That is, we can specify a hypothesis in terms of two locations i and j, and ask whether the causal influence exists from one to the other (due to the time invariance assumption, we are evaluating the link at all time t or for any given time t equivalently).

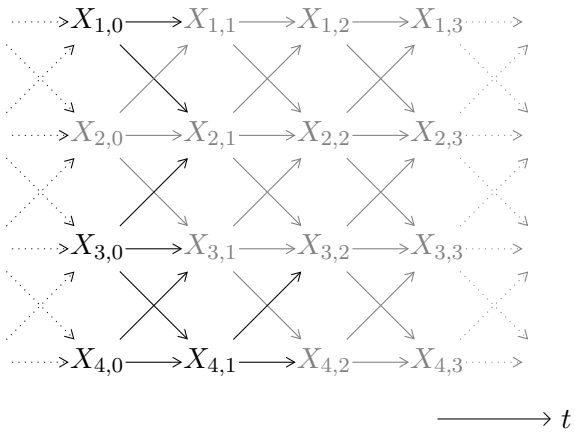


Figure 4.4: "Dynamic" causal model including a light cone shown in grey emanating from $X_{2,0}$, which highlights those nodes (directly or indirectly) causally influenced by $X_{2,0}$.

4.5.1 Ergodicity assumption

This section describes an example that fits the above paradigm for temporal causal models as an illustration of the importance of ergodicity in applying information statistics to time series. A similar system was studied by Ay and Polani (2008) in respect of information flow.

For this example we study two different systems, each comprising two binary time series variables A_t and B_t (that take either 0 or 1). In system 1 the values of the variables are swapped at each time step:

$$\langle a_t, b_t \rangle \leftarrow \langle b_{t-1}, a_{t-1} \rangle$$

On the other hand, the states are inverted at each step in system 2:

$$\langle a_t, b_t \rangle \leftarrow \langle 1 - a_{t-1}, 1 - b_{t-1} \rangle$$

Treating the full state vector of the system at each time step as the state of a Markov chain, the systems are shown graphically in Figure 4.5.

The causal influences in the two systems are quite different. In system 1, node A influences node B and vice versa at each time step. In system 2, the two nodes only influence themselves. Suppose then that we wish to estimate the causal influence of node A on node B, and we construct the model shown in Figure 4.6, which satisfies all of the criteria of our time series modelling approach. The hypothesised link is $H = (A_{t-1}, B_t)$ at any time t (recall the time invariance requirement). It is clear from Figure 4.6 that B_{t-1} d-separates A_{t-1} from B_t. Therefore, we argue that $I(A_{t-1}; B_t | B_{t-1})$ is a suitable test statistic for the hypothesised causal influence H. The information transfer applied to time series in this way is the transfer entropy as defined by

(a) System 1 (b) System 2

Figure 4.5: Example systems in Markov chain state diagram form (in this example all transitions have probability 1).

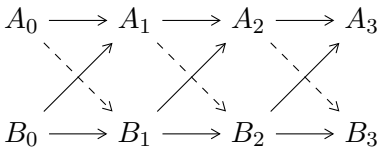


Figure 4.6: Causal time series model of a simple two element system

Schreiber (2000), which we will write as $T_{A\to B}$. This is a non-linear generalisation of "Granger causality" – the notion of causality as the improvement in prediction about the future of the target time series when the past of the causal variable is considered (Barnett, 2009).

Suppose we initialise both systems 1 and 2 at $t = 0$ with the state vector $\langle a_0, b_0\rangle = \langle 1, 0\rangle$. Then in either system, the next state will be $\langle 0, 1\rangle$, and $\langle 1, 0\rangle$ follows from that leading to a period 2 sequence. This sequence is the same in either system and is represented by the horizontal two state subgraph linking $\langle 0, 1\rangle$ and $\langle 1, 0\rangle$ in Figures 4.5a and 4.5b.

We then attempt to estimate the transfer entropy $T_{A\to B}$ by taking the relative frequencies of the states over all time steps in the observed sequence. It is easy to calculate that *for either system* the estimated transfer entropy will be zero. This is problematic, since we know that in one case (system 1) a causal link is present from A to B and in the other (system 2) it is not.

However, this method assumes that the relative frequency of states in the time series is a suitable estimator for the non-interventional probability distribution of those states. But from Figure 4.5 we see that the Markov chains under study are *reducible* and *periodic*,[4] either of which is a sufficient condition for the system to be *non-ergodic*.

The assumption of ergodicity is often used in information theory and time series analysis to justify the use of time-based averages as estimates of probability distributions. In ergodic systems,

[4]Reducible because there are pairs of states (such as $\langle 0, 1\rangle$ and $\langle 1, 1\rangle$) where we cannot transition from one to the other in a finite number of steps. Periodic because all return paths from, e.g. the state $\langle 1, 0\rangle$ back to itself pass through a multiple of 2 transitions.

probabilities estimated using frequencies in time series observations will tend towards the true probability distributions in the long run (irrespective of the starting ensemble distribution – ergodic systems "forget" their initial conditions). A readable justification for this approach can be found in Breiman (1969). Ergodicity was already briefly discussed in chapter 2 as a problem for the causal Markov condition – recall the argument made by Sober (1984, 2001) that the price of bread and the sea level in Venice both rise concurrently but do not have a causal connection. The problem there, as here, is that the data points cannot be translated into probabilities from a single time series without the ergodicity assumption.

If we use the *correct* non-interventional distributions, $P(b_t|a_{t-1}, b_{t-1})$ etc. (which we know from the model equations) we find that the transfer entropy for system 1 is in fact under-determined by the problem as specified. We must also know the probability distribution for the starting state in order to calculate it, that is, we must know the distribution $P(a_0, b_0)$ from which the later states follow (we discuss the precise result further as an appendix to this chapter in section 4.7). However, this starting distribution cannot be established from looking only at a single time series, since we have seen only one instance of $\langle A_0, B_0 \rangle$.

Thus, although it perhaps appears that the transfer entropy $T_{A\to B}$ is zero system 1, contradicting the intuition that it should be positive since there is a genuine causal influence from A to B, the true transfer entropy for this system may in fact take a positive value depending on the distribution $P(a_0, b_0)$. The problem is not that the transfer entropy gives the "wrong" result, but that the non-interventional distributions cannot be sensibly estimated from time series data taken from observation of a non-ergodic system.

Interestingly this problem does not really go away when the information *flow* is used. A result for information flow can be given (see Ay and Polani, 2008) for the case where we have a certain starting distribution, however it is a significant subtlety that this starting distribution is still required for information flow, even though information flow and transfer entropy give different precise values. Indeed, depending on the starting distribution, both information flow and transfer entropy can give a result of zero (again see section 4.7 for a full discussion of both the transfer entropy and information flow results).

4.5.2 Higher dimensions, continuous time series and delay embedding

The previous example had only two variables in the causal network, What about the more general case? What happens, for example, when we increase the number of dimensions? Here we consider how transfer entropy may be somewhat justified for low-dimensional dynamical systems with more than two dimensions. We also investigate the relationship between the transfer entropy and

the physical strength of coupling in dynamical systems.

Figure 4.4 shows a system with 4 dimensions. If we were looking at the hypothesis $H = (X_{1,t-1}, X_{2,t})$ in this context, clearly $X_{2,t-1}$ no longer satisfies the d-separation requirement for information transfer. One possibility is to extend the conditioning set of the transfer entropy so that we condition on *all* the parents of the target variable (in this case X_2) as defined by the topological constraint. For example we could use:

$$I(X_{1,t-1}; X_{2,t} | X_{2,t-1}, X_{3,t-1})$$

Alternatively, transfer entropy is often calculated on a more complete history of the conditioning variable. That is, take the conditional mutual information or transfer entropy with a k-history of the target variable as the conditioning term: $\hat{T}^{(k)}_{1\to 2} = \hat{I}(X_{1,t-1}; X_{2,t} | \mathbf{X}^{(k)}_{2,t-1})$ where $\mathbf{X}^{(k)}_{2,t-1} = \{X_{2,t-k}, \ldots, X_{2,t-1}\}$.

Of course, $\mathbf{X}^{(k)}_{2,t-1}$ still does not strictly d-separate $X_{1,t-1}$ from $X_{2,t}$, since for any finite k there is a path going from $X_{1,t-1}$ to $X_{2,t}$ passing though $X_{2,t-k-1}$ that is not blocked by $\mathbf{X}^{(k)}_{2,t-1}$. However, it is generally accepted that larger values of k give better estimates of the transfer entropy.

A (non-rigorous) way to justify this approach is by analogy to the common practice of attractor reconstruction (Sauer, 2006) in time series analysis. This technique is only applicable to time series generated from continuous time systems. For example, take the well known Lorenz system, defined on a three dimensional state space $\mathbf{v} = (x, y, z)$ with the differential equations:

$$\dot{x} = f_x(\mathbf{v}) = \sigma(y - x) \qquad \dot{y} = f_y(\mathbf{v}) = x(\rho - z) - y \qquad \dot{z} = f_z(\mathbf{v}) = xy - \beta z$$

The parameters σ, ρ and β can be adjusted to create different dynamics, with the combination $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ creating a chaotic, ergodic dynamic (justifying the use of probabilities based on time series). Since the output of the system is continuous in time, we can choose a short time delay τ and after each delay period sample the values of the system variables to generate the time series data. The choice of τ can be made heuristically (Kantz and Schreiber, 2003), for the examples here we use $\tau = 0.2$, approximately one quarter of a cycle.

The system equations suggest a causal model shown in Figure 4.7. The topological constraint is defined by the dynamical functions f – the edge (U_{t-1}, U_t) is present for any variable U, and the edge (W_{t-1}, U_t) is added where W and U are distinct only if w appears in the formula for f_u.

For example, Y_t has the direct parent set of $\{X_{t-1}, Y_{t-1}, Z_{t-1}\}$ because x, y and z all appear in the function f_y (though Y_{t-1} would be included anyway).

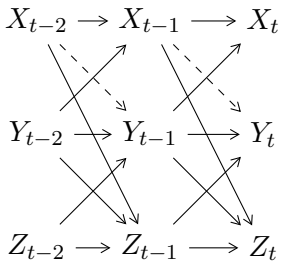


Figure 4.7: Causal model of the Lorenz system, the edge (X_{t-1}, Y_t) has been made "hypothetical" for illustration.

This approach is inevitably somewhat of an approximation. In the case of ordinary differential equations, it simply matches the intuition that intervening on one variable at a particular point in time will immediately impact (in the next short time step) just those variables which depend on it in their dynamical equations. If we solved the system using Euler integration with a time step τ, we would have the generating process (for the Lorenz system):

$$\begin{aligned} x_t &\leftarrow x_{t-1} + \tau\sigma(y_{t-1} - x_{t-1}) \\ y_t &\leftarrow y_{t-1} + \tau(x_{t-1}(\rho - z_{t-1}) - y_{t-1}) \\ z_t &\leftarrow z_{t-1} + \tau(x_{t-1}y_{t-1} - \beta z_{t-1}) \end{aligned}$$

In which case the topological constraint matches literally the generating system. More generally, we can at least be sure that if the differential equations in fact specify two disjoint subsystems (i.e. two uncoupled systems where the variables specific to each subsystem never influence, even indirectly, the variables in the other), then those subsystems will be disjoint in the causal model.

Trajectories in chaotic systems tend towards "strange attractors." Interestingly, and usefully, Takens' delay embedding theorem states that trajectories or attractors in the x, y, z plane can be "reconstructed" by taking a history of length k from any one of the three variables after the temporal sampling, provided that $k > 2m$ where m is the box-counting dimension of the attractor (Sauer and Yorke, 1991; Kantz and Schreiber, 2003). In light of this, consider the transfer entropy $T^{(k)}_{X\to Y} = I(X_{t-1}; Y_t|\mathbf{Y}^{(k)}_{t-1})$ – if k is large, then state of the history $\mathbf{Y}^{(k)}_{t-1} = \{Y_{t-k}, Y_{t-k}, \ldots, Y_{t-1}\}$ should be enough to determine the full state of the system at time $t-k$, i.e. $\{X_{t-k}, Y_{t-k}, Z_{t-k}\}$. This *is* a state which d-separates X_{t-1} from Y_t according to the causal graph.

In realistic scenarios, the embedding theorem may not apply exactly (in the case that the system under study is purely a low-dimensional dynamical system). Thus the justification of transfer entropy as a statistic for causal inference in general depends on a heuristic assumption that the

history (e.g. $\mathbf{Y}_{t-1}^{(k)}$) gets "close to" d-separating the causal (X_{t-1}) and effect (Y_t) variables by virtue of containing information about the full state of the system at a prior time point.

As an illustration, consider the null hypothesis of no causal interaction in a pair of coupled Lorenz systems. Each system is denoted by the subscript index $i \in \{1, 2\}$, and we rewrite the Lorenz equations as:

$$\dot{x}_i = \sigma(y_i' - x_i) \qquad \dot{y}_i = x_i(\rho - z_i) - y_i' \qquad \dot{z}_i = x_i y_i' - \beta z_i$$

The y_i' are introduced to allow coupling the two systems – we introduce a coupling parameter $\gamma \in [0, 1]$ and set:

$$y_1' = (1 - \gamma)y_1 + \gamma y_2 \qquad y_2' = (1 - \gamma)y_1 + \gamma y_2$$

So for $\gamma = 0$ we have two fully independent Lorenz systems, but as γ is increased the systems become increasingly coupled. At around $\gamma = 0.6$ a transition is observed to a synchronised chaotic dynamic – both systems remain chaotic but each follows the trajectory of the other exactly. This is an example of chaotic synchronisation, comparable to the synchrony experiment in chapter 3 (see also Pikovsky et al., 2001).

For non-zero coupling, the two systems causally interact through the y variable, and we therefore expect positive transfer entropy $\hat{T}_{Y_1 \to Y_2}$. The estimated transfer entropy is shown in Figure 4.8 as the solid thick line, and we see that the estimated transfer entropy initially increases for low coupling values, but for very high coupling, when the systems become fully synchronised, the transfer entropy returns to zero. This results from the lack of ergodicity that occurs when the system becomes synchronised – since the Y_1 and Y_2 series are exactly the same in this case, any history of $\mathbf{Y}_{2.t-1}^{(k)}$ is bound to condition out the mutual information between $Y_{1,t-1}$ and $Y_{2,t}$.

This exact synchrony, leading to loss of ergodicity, changes if we add noise into the system. A stochastic simulation can be obtained using the order 1.0 Runge-Kutta integration method for stochastic differential equations (Sauer, 2012). The grey lines in Figure 4.8 show how the transfer entropy increases in the strong synchrony region when even a small amount of noise is added to the system (the maximum noise variance we used was $\sigma^2 = 1$, which is small compared to the cycle amplitude of the Lorenz system). The stochastic element causes enough variation in the sequence that the causal influence can be detected. Note that this stochastic variation does not entirely destroy the synchronisation effect – the coupled systems are still closely synchronised, as can be seen from the time series in Figure 4.9. This means that even when the system is ergodic

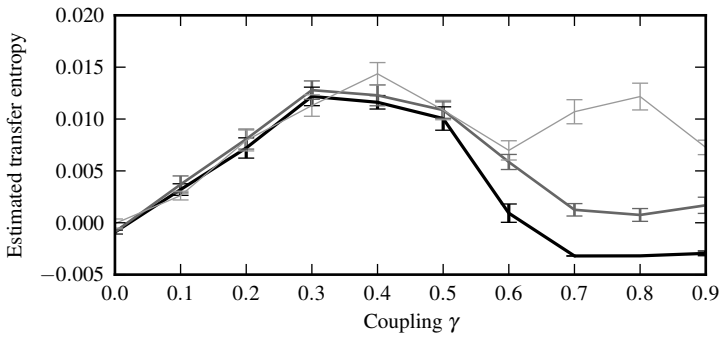


Figure 4.8: Transfer entropy for coupled Lorenz system. Lines and error bars show mean and standard error of 10 samples of $\hat{I}(Y_{1,t-1}; Y_{2,t}|\mathbf{Y}^{(4)}_{2,t-1}) - \frac{16}{2n}$ where $\hat{I}$ is obtained from a binary maximum entropy binning of all the variables on a simulated coupled Lorenz system, and $\frac{16}{2n}$ is the bias of the statistic with 16 degrees of freedom and $n = 2496$ samples. The thick back line has no dynamical noise. The other two lines are estimated with simulated Gaussian noise of standard deviation $\sigma^2 = 0.1$ (grey) and 1 (light grey).

(after introducing a small amount of noise), the transfer entropy can decrease even as the coupling strength γ increases.

4.6 Discussion

This chapter has approached two seemingly distinct issues and it is important to draw them together at this point. First, we have seen that information flow does not necessarily provide an improvement over information transfer for inferring the presence of causal influence. Second, it has been argued that we must distinguish between the inferential value of a statistic and the measurement of *causal strength*.

In fact, these are two sides of the same problem. Information transfer only appears to have problems as a tool for inference as long as we are *expecting* it to also measure strength. In the coupled Lorenz system, for example, we saw that increasing coupling strength first led to an increase in transfer entropy, but then transfer entropy drops under strong coupling. This leads us to naively interpret transfer entropy as having failed as a tool for detecting causality, and encourages looking for other statistics such as information flow that would be better at detecting causality.

However, from the inferential perspective, we can see that transfer entropy is unlikely to be positive in the absence of a causal influence, and so high transfer entropy values do have inferential utility, according to the theory of inference discussed in section 4.2. The problem is only that low transfer entropy may occur *both* due to no causal influence, or because of a very strong causal influence.

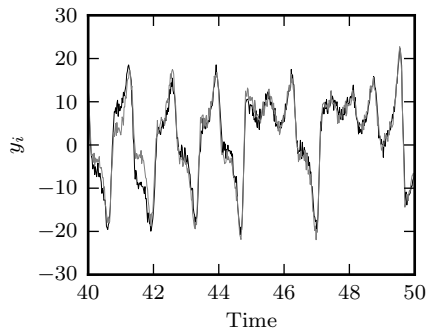


Figure 4.9: A short time period of strongly coupled Lorenz systems. The coupling parameter is $\gamma = 0.7$ so the Lorenz systems are highly synchronised. As a result the two variables $y_i, i \in \{1, 2\}$ follow each other closely. A relatively large Gaussian noise component with standard deviation $\sigma^2 = 1$ is added, so the two time series are seen to deviate slightly from each other, but the general synchronisation is maintained.

This asymmetry is highlighted by the Neyman-Pearson model of testing introduced at the beginning of this chapter. Recall that the statistical power $1 - \beta$ is defined as the probability of a positive result when the positive hypothesis is in fact true. The example in 4.4.1 showed calculations of the power for information statistics under a specific positive hypothesis (the presence of a defined causal influence). The statistical power has two important and related ramifications:

1. Literally, it tells us the probability of detecting a causal influence when that causal influence is present – since it monotonically increases with the exact value of the information under the positive hypothesis that the causal influence is indeed present, the information is a proxy for statistical power. That is, the information is a measure, not of the strength of a causal influence but of the *epistemological* "obviousness" or clarity of that causal influence, under a particular observational (information transfer) or interventional (information flow) regime.

2. Moreover, the power (and by association information) places an upper limit on the *severity* (as defined in section 4.2) with which we can reject the presence of that causal influence in the case where we measure a low empirical information transfer or flow. That is, when we have a strong enough theory to specify the type of causal influence we are hypothesising, we can calculate the power and hence the inferential value of a "negative" (low information) result.

Information flow is a useful inferential statistic, but no more so than information transfer. We cannot say in the general case that one is preferable over the other – which is more useful from an inferential perspective is a consequence of the power and severity considerations described.

This chapter demonstrates the causal interpretation of information transfer, and the related time series statistic transfer entropy. The causal graphical modelling approach of Pearl (2009) (similar to Spirtes et al., 2001) has been adopted to make this interpretation clearer.

In time series, I adopt a structural, mechanical view of causality where causal influence involves the transfer of information at a limited velocity from a point where the information is introduced at X_t to a point where its effect is seen at Y_{t+n} (cf. the "mark transmission theory of Salmon, 1984, 1998). "Chain DAG" temporal models as used here are found in various forms in the literature, but are not the only approach to causality in time series, for example, some techniques allow causal influences to cross multiple time lags or to occur "instantaneously," resulting in a slightly different conception of temporal causality, effectively relaxing the temporal Markov constraint used here (Rissanen, 1987; Eichler, 2012a,b; Amblard and Michel, 2011; Runge et al., 2012b). Though the these approaches are not immediately commensurate with the one presented here, there are clearly a number of parallels in the conceptions used, in particular, the use of Granger-type notions of temporal causality.

Particular emphasis is placed here on the ergodicity assumption, which justifies interpreting the proportion of time the system spends in a given state as the probability of that state at any given time. We do not have to look hard for precedents for this assumption in the information theory literature: Shannon (1948) and Wiener (1965) gave it central importance, and further motivations for its importance can be found in modern references on time series such analysis (Breiman, 1969; Kantz and Schreiber, 2003). However, it seems to have been recently neglected somewhat: Ay and Polani (2008) give a non-ergodic example to show the limits of information transfer, but do not mention the word ergodicity. Eichler (2012a) states that (emphasis added): "because of *stationarity*, edges in this graph [referring to temporal DAGs of the form used in the current paper] are translation invariant," justifying the equivalence of an edge from $X_{1,t}$ to $X_{2,t+1}$ at a given time t to the same edge at all times t. This is correct in a certain sense, but it glosses over the significant difference between stationarity and ergodicity.[5] Without ergodicity, even the interventionally constructed information flow does not define a unique quantity when applied to time series.

Violations of ergodicity are likely to lead to underestimation of causal influence (overestimation is not possible since, even without ergodicity, d-separation in the causal structure implies statistical independence in the data). Here a simple synchronising system has been used to demonstrate this, a similar but more complex example is discussed in chapter 3. The ergodicity problem

[5] In fact, Eichler (2012a) does assume ergodicity, which can be obtained in Gaussian autoregressive models from conditions on the covariance matrix or the spectral density matrix. Assumption 2 in that reference adds such a condition, but does not call it "ergodicity."

can naturally also be avoided when a system is simulated from a known distribution of starting conditions, and estimates obtained by averaging across the ensemble (e.g. Williams and Beer, 2010a; Lizier and Prokopenko, 2010; Ceguerra et al., 2011).

4.6.1 Information dynamics and complexity

A final point about information has not been discussed above but should not be neglected as it is central to the arguments of much of the work cited. An outgrowth of the modern approach to complexity science is the view that almost everything can be described *computationally* – in terms of information transfer, storage and processing (e.g. Crutchfield and Mitchell, 1995; Mitchell, 2006; Lizier, 2010). This view has its own motivations, not necessarily related to establishing causal effects. Modern scientific models are quite often literally computer programs. Even when they are not, it seems natural to ask, for example, what information sensory organs provide to guide an organism's behaviour or how a school of fish "communicates" to form a coherent swarm. Information is transferred between components, stored at various locations, old information is processed new information is "computed" by the functioning and interaction of system components. Computational and information theoretic views of this sort are increasingly found in the study of, for example, biological systems (Kitano, 2002; Cohen, 2006; Nurse, 2008), human, animal or robot behaviour (Pfeifer et al., 2007b; Lungarella and Sporns, 2006; Klyubin et al., 2008; Pitti et al., 2009), physical or environmental dynamics (Baptista and Kurths, 2005; Runge et al., 2012b) and economic and social mechanisms (Epstein, 1999; Rosvall and Bergstrom, 2007; Oka and Ikegami, 2012).

This focus on information as a statistical inference tool largely ignores the non-linear properties of information, namely that unlike linear correlation, the mutual information increases as the relationship between the measured variables becomes more "complex." This view leads to a number of information theoretic definitions of complexity – some examples being Bialek et al. (2001); Crutchfield and Young (1989) and Rissanen (1986).

Information flow is derived from calculations of causal effects, and thus explicitly depends on the causal model. Lizier and Prokopenko (2010) argue that only information flow reflects causality and information transfer should be used to describe computation, which we might regard as being a "complexity" view of information transfer. In fact we can correctly view both information flow and information transfer as measures of *both* causal influence *and* complexity (with the caveat that information transfer must be defined with respect to causal criteria as we have done here). In the case of causal influence, we can say that the difference between the two is a question of inferential power. In the case of complexity there is clearly a difference (since the two values do

not always coincide), but the exact nature of this difference is perhaps unclear as a result of the inherent ambiguity of the word "complexity."

Hopefully, this analysis puts on a clearer footing the understanding of information transfer as a measure of causal influence that is often alluded to but not explicated in the literature. On the one hand, it is clearly appropriate to be cautious about inferring causal influences from observational data, but on the other, we need a good understanding of what observational statistics are actually showing, and to acknowledge that in fact the causal interpretation is not without justification.

4.7 Information flow and transfer entropy for Markov chains

We discuss the general results for the Markov chain "system 1" in section 4.5.1. We show that both transfer entropy and information flow are undefined without knowledge of the initial distribution $P(a_0, b_0)$. We also show that they may be undefined even when we do know this. The results for one particular starting distribution are given for this system by Ay and Polani (2008) – we discuss the more general case where the starting distribution can be anything.

The transfer entropy of interest is $TE_{A\to B}$ defined as

$$I(A_{t-1}; B_t | B_{t-1}) = \sum P(a_{t-1}, b_t, b_{t-1}) \log \frac{P(b_t | a_{t-1}, b_{t-1})}{P(b_t | b_{t-1})}$$

The sum is taken over all possible combinations of values for a_{t-1}, b_t and b_{t-1} (where each can be either 0 or 1). But note that the system definition ensures that b_t must always be the same as a_{t-1}, and so

$$P(a_{t-1}, b_t, b_{t-1}) = \begin{cases} P(a_{t-1}, b_{t-1}) & b_t = a_{t-1} \\ 0 & b_t \neq a_{t-1} \end{cases}$$

We can thus neglect terms from the sum where $b_t \neq a_{t-1}$, and substitute the joint distribution as follows from the above:

$$\sum P(a_{t-1}, b_{t-1}) \log \frac{P(b_t | a_{t-1}, b_{t-1})}{P(b_t | b_{t-1})}$$

Furthermore, we are only considering terms of the sum where $b_t = a_{t-1}$ we can make the following substitutions:

$$P(b_t | a_{t-1}, b_{t-1}) = P(a_{t-1} | a_{t-1}, b_{t-1}) = 1$$

and

$$P(b_t|b_{t-1}) = P(a_{t-1}|b_{t-1})$$

Thus the transfer entropy becomes:

$$\begin{aligned} I(A_{t-1}; B_t|B_{t-1}) &= \sum P(a_{t-1}, b_{t-1}) \log \frac{P(b_t|a_{t-1}, b_{t-1}}{P(b_t|b_{t-1})} \\ &= \sum P(a_{t-1}, b_{t-1}) \log \frac{1}{P(a_{t-1}|b_{t-1})} \\ &= H(A_{t-1}|B_{t-1}) \end{aligned}$$

This value is not defined by the problem definition as specified, unless we give a distribution at the start time $t = 0$ from which the initial values a_0 and b_0 are drawn:

$$P(a_0, b_0)$$

From which we can arrive at $H(A_0|B_0) = \sum P(a_0, b_0) \log \frac{P(b_0)}{P(a_0,b_0)}$. Since the variables are swapped at each time step, it follows that $H(B_1|A_1) = H(A_0|B_0)$. In general $H(B_t|A_t) = H(A_{t-1}|B_{t-1})$, by the symmetry of the system also $H(A_t|B_t) = H(B_{t-1}|A_{t-1}) = H(A_{t-2}|B_{t-2})$.

Inductively, it follows that

$$\begin{aligned} I(A_{t-1}; B_t|B_{t-1}) &= H(A_{t-1}|B_{t-1}) \\ &= \begin{cases} H(A_0|B_0) & t \text{ even} \\ H(B_0|A_0) & t \text{ odd} \end{cases} \end{aligned}$$

This gives the peculiar result that the transfer entropy is dependent on the actual value of t referred to. Usually, we would consider t a "dummy" variable – since typically the transfer entropy should be invariant to the choice of t. This will happen in cases where $H(A_0|B_0) = H(B_0|A_0)$, which is true at least for the starting distribution given by Ay and Polani (2008), namely

$$P(a_0, b_0) = \begin{cases} 0.5 & a_0 \neq b_0 \\ 0 & \text{otherwise} \end{cases}$$

That is, the starting states $\langle 0, 1\rangle$ and $\langle 1, 0\rangle$ have equal probability, and other starting states ($\langle 0, 0\rangle$ and $\langle 1, 1\rangle$) can never occur. In this case, the transfer entropy is zero as noted by Ay and Polani (2008), since $H(A_0|B_0) = 0$ – this follows from the fact that the value of A_0 is always 1 if B_0 is 0, and 0 if B_0 is 1.

However there are other starting distributions that give defined results for transfer entropy that are non-zero: for example allowing all four start states with equal probability. In this case, $H(A_0|B_0) = 1$ bit, and so transfer entropy is 1 bit, corresponding to the intuition that A is a cause of B in system 1.

For the equivalent information flow, recall the formula:

$$IF(A_{t-1} \to B_t|\widetilde{B}_{t-1}) = \sum_{a_{t-1},b_t,b_{t-1}} P(b_{t-1})P(a_{t-1}|\widetilde{b}_{t-1})P(b_t|\widetilde{a}_{t-1},\widetilde{b}_{t-1})\times \log \frac{P(b_t|\widetilde{a}_{t-1},\widetilde{b}_{t-1})}{\sum_{a'_{t-1}} P(a'_{t-1}|\widetilde{b}_{t-1})P(b_t|\widetilde{a}'_{t-1},\widetilde{b}_{t-1})}$$

We can again neglect terms in the main sum, since $P(b_t|\widetilde{a}_{t-1},\widetilde{b}_{t-1})$ is clearly 1 if $b_t = \widetilde{a}_{t-1}$ and 0 otherwise. We can also neglect all but one term of the sum in the denominator of the fraction inside the log: $P(b_t|\widetilde{a}'_{t-1},\widetilde{b}_{t-1})$ is 1 if and only if $a'_{t-1} = b_t = a_{t-1}$. Furthermore, notice that $P(a_{t-1}|\widetilde{b}_{t-1}) = P(a_{t-1})$ as a result of the intervention on b_{t-1} – this intervention ensures that b_{t-1} is not dependent on any causal antecedent it might share with a_{t-1}, and since the two variables have the same time index clearly b_{t-1} cannot influence the value of a_{t-1}.

$$\begin{aligned}
IF(A_{t-1} \to B_t|\widetilde{B}_{t-1}) &= \textstyle\sum_{a_{t-1},b_{t-1}} P(b_{t-1})P(a_{t-1}|\widetilde{b}_{t-1}) \log \frac{1}{P(a_{t-1}|\widetilde{b}_{t-1})} \\
&= \textstyle\sum_{a_{t-1},b_{t-1}} P(b_{t-1})P(a_{t-1}) \log \frac{1}{P(a_{t-1})} \\
&= \textstyle\sum_{b_{t-1}} P(b_{t-1}) \sum_{a_{t-1}} P(a_{t-1}) \log \frac{1}{P(a_{t-1})} \\
&= \textstyle\sum_{a_{t-1}} P(a_{t-1}) \log \frac{1}{P(a_{t-1})} \\
&= H(A_{t-1}) \\
&= \begin{cases} H(A_0) & t \text{ even} \\ H(B_0) & t \text{ odd} \end{cases}
\end{aligned}$$

following a similar inductive argument as before. Hence the information flow is again undefined in the general case. For starting distribution with $\langle 0,1\rangle$ and $\langle 1,0\rangle$ having probability 0.5 each, it is however 1 bit. Of course it could also be zero, for example if the starting distribution allowed only one possible configuration with probability one.

The key point from the above is that neither the information flow nor the transfer entropy necessarily give a result that corresponds directly to the intuition about cause: we know in system 1 that A causes B, but depending on the distribution for the starting state, both the transfer entropy and information flow can be zero.

We should also note that some of the starting distributions given do not correspond to our intuitions about causal DAGs. If we consider our general causal DAG time series model for two variables A and B in figure 4.6, note that we have drawn A_0 and B_0 as separate nodes, which is natural according to our definition of a temporal DAG. However, for the CMC to hold with respect to this DAG, note that the two "root" nodes of the graph (the nodes with no parents – A_0 and B_0) should be statistically independent of each other (recall this is the Markovian assumption from with Pearl (2009) derives the CMC, see chapter 2).

Some of the starting distributions we have considered, for example the one used by Ay and Polani (2008) which allows only $\langle 0, 1\rangle$ and $\langle 1, 0\rangle$ as starting states clearly do not conform to this requirement. Ay and Polani (2008) avoid having a non-Markovian graph by making a single starting node for the combination A_0, B_0.

However, if we wish to maintain our graphical model where A_0 and B_0 are separate nodes, then if the system is Markovian with respect to this graph, A_0 and B_0 must be statistically independent, i.e. the starting distribution must satisfy

$$P(a_0, b_0) = P(a_0)P(b_0)$$

It follows from this that $H(A_0|B_0) = H(A_0)$ and hence for systems which are Markovian in this respect, the information flow and transfer entropy must be the same.

Chapter 5

Convergent cross-mapping and transfer entropy

This chapter discusses the convergent cross-mapping (CCM) technique for detecting causal influences. This technique has been advocated recently as being particularly suitable to the study of non-linear dynamical systems, being based on a substantively different concept of causation than the Granger type methods (including transfer entropy) that we adopt elsewhere in this thesis (Sugihara et al., 2012).

This chapter attempts to place CCM on a similar footing to the other work in this thesis, by deriving an information theoretic analogue of CCM, which we call cross-embedded mutual information (CMI). Doing so allows us to make a comparison between CCM and transfer entropy by placing them both in the same information theoretic framework in the formulation of CMI.

Furthermore we argue that CMI has a number of theoretical and practical advantages over CCM, but under fairly mild conditions CMI approximates CCM. Furthermore CMI is bounded from below by the time-delayed mutual information (TDMI). As a result, CCM and CMI have many of the same disadvantages as TDMI relative to the currently more widespread approach of transfer entropy. Specifically, while transfer entropy is known to be vulnerable to non-separability of antecedent predictors, CCM/CMI is vulnerable to non-causal information transfer (like TDMI). This suggests that the difference between the new CCM method and more established Granger causal methods including TE is largely a sensitivity/specificity trade-off (i.e. we trade false negatives for false positives), rather than as previously claimed a consistent advantage for CCM for analysis of coupled non-linear dynamical systems.

5.1 Introduction

Convergent cross-mapping (CCM) (Sugihara et al., 2012) has recently been proposed as a novel method for detecting causal relationships in weakly coupled non-linear dynamical systems. Primarily the approach consists of assessing the statistical correlation between a putative causal time series variable X and a prediction of X derived from a delay embedding of a potential effect Y. Sugihara et al. (2012) present CCM in contradistinction to the more established Granger causality paradigm (Granger, 1969), which starts from different assumptions about the nature of causation. Granger causality and related methods such as transfer entropy (Schreiber, 2000) (generally seen as a non-linear generalisation of the more traditional Gaussian auto-regression approach to Granger causality (Barnett, 2009)) represent causal relationships as the relative improvement in the prediction of the effect variable when the state of the causal variable is added to the pre-existing information about the state of the world excluding the causal variable.

The advantages claimed by Sugihara et al. (2012) for CCM over Granger causality are in dealing with non-linearity and non-separability of the causal variable. Non-linearity is likely to be less of a problem for transfer entropy, which is inherently a model-free statistic (by virtue of its information theoretic formulation), than auto-regression based Granger causality.

Separability refers to a well known problem with Granger causality (in fact noted by Granger (1969)) – that it assumes it will be possible to exclude, or separate, genuine causes from common causes. That is, Granger causality measures the difference in uncertainty regarding the "effect" variable between two states of background knowledge – in one, we have access to all prior information including the causal variable, and in the other we have access to all prior information excluding the causal variable. As a result we must assume that the available measurements of variables that are not the putative cause (e.g. the historical state of the effect variable) do not contain information about the state of the causal variable. This is generally not the case for coupled dynamical systems, and can lead to non-detection of genuine causal influences. This problem applies equally to transfer entropy.

This chapter investigates the claimed advantages of CCM over Granger casual methods including TE. Section 5.2 reviews CCM and attempts to formalise the sense of "causation" that appears to be intended by Sugihara et al. (2012), in effect we assume causation in this context means non-zero (and non-trivial) coupling between dynamical systems. Section 5.3 introduces our proposed alternative to CCM, cross-embedded mutual information (CMI)

$$CMI_{X \to Y} = I(X_t; Y_t, Y_{t+1}, \ldots, Y_{t+m-1}) \tag{5.1}$$

and discusses conditions under which CMI is bounded from below by a monotonic isomorph-

ism of CCM.

Reformulation as CMI makes clear the distinction (and similarities) between CCM and Granger methods. However, as discussed in section 5.4, CMI is bounded from below by time delayed mutual information (TDMI), which is known to overestimate causal influences as a result of non-causal information transfer (correlations introduced by common causes). Section 5.5 compares CCM and CMI with two transfer entropy measures. This appears to confirm first that CMI approximates CCM, and that CCM/CMI trades spurious non-detection of genuine causes in Granger/TE methods (due to non-separability) for spurious over-detection of non-existent causes (due to non-causal information transfer).

5.2 Convergent cross-mapping

Consider a two variable discrete time dynamical system defined as follows:

$$\begin{aligned} x_{t+1} &= f_x(x_t, y_t) \\ y_{t+1} &= f_y(x_t, y_t) \end{aligned} \tag{5.2}$$

With $t \in \mathbb{Z}_+$ and $(x_t, y_t) \in (\mathcal{X}, \mathcal{Y}) \subseteq \mathbb{R}^2$. Without subscripts variables represent complete time series, e.g. $x = \{x_0, x_1, ...\}$. The initial values x_0, y_0 are a sampled from specified random variables X_0, Y_0. For each subsequent t, x_t, y_t can be likewise treated as samples from X_t, Y_t, particularly if we assume the system is chaotic. Further assume the system is stationary ergodic for validity of time-series estimates, violation of such assumptions are known to produce problems for causal inference (Hoover, 2003).

The following arguments are likely to apply to continuous time systems with appropriate modifications, but we consider only discrete time systems.

Say that "X causes Y" if it is possible (in probability, with respect to the stationary measures of X_t and Y_t) that a different value of x_t might alter the function f_y:

$$Cause(X, Y) \Leftrightarrow \neg \left[f_y(x_t, y_t) \overset{a.s.}{=} f_y(x'_t, y_t) \right] \tag{5.3}$$

Otherwise, "X does not cause Y", and conversely for Y causes / does not cause X. For example, in the pair of coupled logistic map functions:

$$\begin{aligned} x_{t+1} &= qx'_t(1-x'_t) \\ y_{t+1} &= qy'_t(1-y'_t) \\ x'_t &= (1-\beta_{xy})x_t + \beta_{xy}y_t \\ y'_t &= (1-\beta_{yx})y_t + \beta_{yx}x_t \\ x_0 &\in (0,1) \\ y_0 &\in (0,1) \end{aligned} \tag{5.4}$$

Where q is a parameter in $(0, 4)$ and β_{xy} and β_{yx} represent coupling parameters in $[0, 1)$, we find that X causes Y if and only if $\beta_{yx} > 0$, and conversely $\beta_{yx} = 0$ means X does not cause Y.

This definition of causation is intended to follow the one implied by Sugihara et al. (2012), and is deliberately distinct from the one usually adopted in discussions of Granger causality, wherein a cause precedes an effect, and causes and effects are statistically independent after conditioning on all common causes (Granger, 1969; Holland, 1986). Our definition is motivated by attempting to capture the idea that different "possibilities" for the state of a causal variable should lead to different outcomes for the state of the effect variable.

However, we should remark the our definition of causation does bear some resemblance to the structural equation models (SEM) used by Pearl (2009) and already seen in the previous chapter (section 4.3 – in which causes are similarly defined as arguments to functions determining their effects. The most obvious difference is that SEM models usually posit independent noise added to each variable on top of the functional relationship, such that the model is irreducibly stochastic (whereas our models are deterministic, if often chaotic).

The key theory underlying CCM is that if X causes Y in the above sense, then an m-dimensional delay embedding of y defined as

$$\begin{aligned} y^{(m)} &= \{y_0^{(m)}, y_1^{(m)}, \ldots\} \\ y_t^{(m)} &= \{y_t, y_{t+1}, \ldots, y_{t+m-1}\} \end{aligned}$$

is diffeomorphic to the combined dynamical system (x, y) provided the dynamics of the system are close to an attractor and m is sufficiently large (greater than twice the dimension of the attractor) according to the Takens delay embedding theorem (Takens, 1981; Sugihara et al., 2012). Thus with knowledge of this diffeomorphism, x_t could be calculated from an observation of $y_t^{(m)}$. This argument can be extended to arbitrary numbers of dynamical variables, x, y, z, etc, where any two are selected to be considered for a direct causal influence.

Of course we do not have exact knowledge of the diffeomorphism (and real-life data would likely be distorted by noise in any case). However, given a data set of observations, $D = \{(x_t, y_t^{(m)}) : t \in T_{obs}\}$, we can use a non-linear regression of x onto $y^{(m)}$ to produce an es-

timator $\hat{x}(y_t^{(m)})$ which can be used to find likely values for x_t given a previously unseen value of $y_t^{(m)}$. If $y_t^{(m)}$ is diffeomorphic to (y_t, x_t), in other words if X causes Y, then this estimator should perform well.

From this line of reasoning, Sugihara et al. (2012) propose to assess causality by evaluating the accuracy of such an estimator. Specifically, we will here define CCM (following Sugihara et al., 2012) as the correlation between the estimated and true values of x_t in the sample D via Pearson's r:

$$\begin{aligned} CCM_{X\to Y} &= r\left(X_t, \hat{x}(Y_t^{(m)})\right)^2 \\ &= \left[\frac{\mathrm{cov}[X_t, \hat{x}(Y_t^{(m)})]}{\mathrm{std}[X_t]\mathrm{std}[\hat{x}(Y_t^{(m)})]}\right]^2 \end{aligned} \tag{5.5}$$

Note that the assumption of ergodicity is employed above since we assume the ergodic theorem, *viz.* the limit in probability

$$\lim_{|D|\to\infty} \mathbb{E}[f(X_t)] \overset{a.s.}{=} \frac{1}{|D|}\sum_{x_t\in D} f(x_t) \tag{5.6}$$

can be used to approximate r^2 (by first approximating the true covariance and standard deviations in equation 5.5) from the data points in an empirical data set D.

High values of CCM (close to unity) thus represent good performance of the estimator and are used to infer the presence of a causal relationship. Note Sugihara et al. (2012) generally report r, rather than r^2 – the choice of r^2 here is to simplify the discussion below. Since the correlation between the estimator and true value will be positive if it exists, this does not lose any information of interest.

In practise we use k nearest neighbour (k-NN) regression to construct $\hat{x}$ and "leave-one-out" cross-validation, following Sugihara et al. (2012), to evaluate r^2 (i.e. for each chosen $y_t^{(m)} \in D$, $\hat{x}(y_t^{(m)})$ is trained with the corresponding data point excluded, i.e. it is constructed from $D \setminus (x_t, y_t^{(m)})$, since if $\hat{x}(y_t^{(m)})$ had access to $(x_t, y_t^{(m)})$, we could trivially ensure $r = 1$ by setting $\hat{x}(y_t^{(m)}) = x_t$ always). Figure 5.1 illustrates CCM in weakly coupled logistic maps.

5.3 Cross-embedded mutual information

This section defines an information theoretic measure, cross-embedded mutual information (CMI), which is comparable to CCM, and shows that under certain conditions we can expect it to be bounded from below by a monotonic function of CCM. We motivate the new measure, CMI,

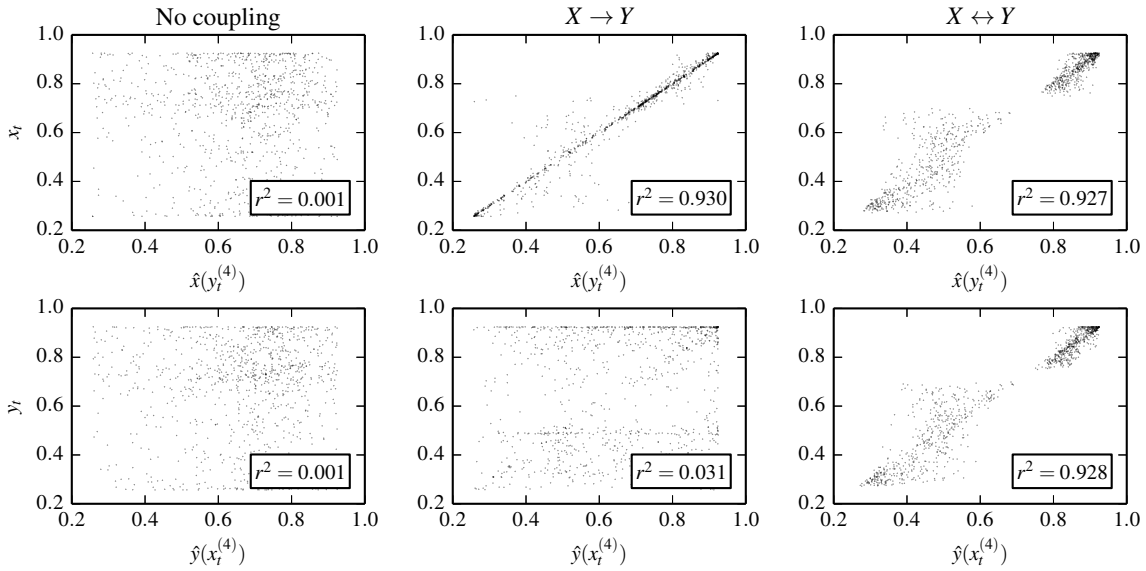


Figure 5.1: Demonstration of CCM using the coupled logistic maps in equation 5.4. The top row shows estimates of x based on time series embedding of y, whereas the bottom row shows estimates of y based on embedding x. The r^2 values shown represent the CCM estimates as described in the text. On the left we have $\beta_{xy} = \beta_{yx} = 0$ – no coupling, and hence poor performance of the estimates. Setting $\beta_{yx} = 0.1$ (X a cause of Y, but $\beta_{xy} = 0$, Y not a cause of X) introduces correlation between x and $\hat{x}$, but not y and $\hat{y}$. Coupling in both directions, $\beta_{xy} = \beta_{yx} = 0.02$ introduces correlation in both directions.

by deriving it from the existing definition of CCM. This will also serve to show the similarities between the two approaches.

Recall that CCM evaluates the performance of the estimator $\hat{x}$ using Pearson's r. Here, we will consider measuring the same performance using mutual information. First define mutual information between two random variables X and Y (see (Cover and Thomas, 2006))

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= h(X) - h(X|Y) \end{aligned} \tag{5.7}$$

where $h(\cdot)$ ($h(\cdot|\cdot)$) is the differential (conditional) Shannon entropy of a continuous random variable

$$\begin{aligned} h(V) &= -\int_{supp(V)} p(v) \ln p(v) \mathrm{d}v \\ h(V|U) &= -\int\int_{supp(V,U)} p(v,u) \ln p(v|u) \mathrm{d}u\mathrm{d}v \end{aligned} \tag{5.8}$$

Also define conditional mutual information between X and Y given Z:

$$I(X;Y|Z) = h(X|Z) - h(X|Y,Z) \tag{5.9}$$

Mutual information has some conceptual similarities to r though they can be arbitrarily unrelated. It is typically viewed as measuring statistical dependence – $I(X;Y) = 0$ if and only if X is statistically independent of Y – $p(X,Y) \equiv p(X)p(Y)$ – and $I(X;Y) > 0$ otherwise. Mutual information is known to have desirable "equitability" properties relative to r (Kinney and Atwal, 2014), meaning that it is generally invariant to non-linearities in the dependence between two variables. For linear relationships, mutual information often increases monotonically with r^2, however, there is no fixed relationship between mutual information and r independent of the marginal distributions of the variables under consideration (Foster and Grassberger, 2011).

However, if we consider a linear relationship of the form $Y = X + \eta$ where η is i.i.d. Gaussian noise

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+\eta|X) \\ &= h(Y) - h(\eta|X) \\ &\geq h(Y) - h(\eta) \end{aligned} \tag{5.10}$$

since $h(\eta) \geq h(\eta|X)$, with equality when η is statistically independent of X. If we consider the Gaussian variable $Y^{\mathcal{N}}$ with first and second moments equal to Y, i.e. $Y^{\mathcal{N}} \sim \mathcal{N}(\mathbb{E}[Y], \mathrm{var}[Y])$, the information theoretic quantity D_{KL} (Kullback-Leibler divergence) is

$$\begin{aligned} D_{KL}(Y||Y^{\mathcal{N}}) &= h_Y(Y^{\mathcal{N}}) - h(Y) \\ &= h(Y^{\mathcal{N}}) - h(Y) \end{aligned} \tag{5.11}$$

where $h_Y(Y^{\mathcal{N}})$ is the cross entropy

$$h_Y(Y^{\mathcal{N}}) = -\int_{supp(Y)} p_Y(y) \log p_{Y^{\mathcal{N}}}(y) \mathrm{d}y \tag{5.12}$$

but since $Y^{\mathcal{N}}$ is Gaussian by construction, $h_Y(Y^{\mathcal{N}}) = h(Y^{\mathcal{N}})$.

Note that considering that all $D_{KL} \geq 0$ one can find from equation 5.11 that $h(Y^{\mathcal{N}}) \geq h(Y)$ and likewise $h(\eta^{\mathcal{N}}) \geq h(\eta)$. This is the well known result that differential entropy is bounded from above by the differential entropy of an equal mean and variance Gaussian distribution (Cover and Thomas, 2006). As a result, deviation from the assumption that η is Gaussian gives

$$h(\eta) \leq -\frac{1}{2} \ln(2\pi e \cdot \mathrm{var}[\eta]) \tag{5.13}$$

Now, substituting 5.11 and 5.13 into 5.10

$$\begin{aligned} I(X;Y) &\geq h(Y^{\mathcal{N}}) - h(\eta) - D_{KL}(Y||Y^{\mathcal{N}}) \\ &\geq \tfrac{1}{2} \ln(2\pi e \cdot \mathrm{var}[Y^{\mathcal{N}}]) \\ &\quad -\tfrac{1}{2} \ln(2\pi e \cdot \mathrm{var}[\eta]) \\ &\quad -D_{KL}(Y||Y^{\mathcal{N}}) \end{aligned} \tag{5.14}$$

By construction, $\mathrm{var}[Y^{\mathcal{N}}] = \mathrm{var}[Y]$ and $\mathrm{var}[\eta] = \mathrm{var}[Y](1 - r(X,Y)^2)$ (recalling that η is the residuals of a linear model), so

$$I(X;Y) \geq -\frac{1}{2} \ln(1 - r(X,Y)^2) - D_{KL}(Y||Y^{\mathcal{N}}) \tag{5.15}$$

Recalling our estimator, $\hat{x}_t = \hat{x}(Y_t^{(m)})$, if we have an approximately linear relationship $X_t = \hat{x}_t + \eta$ (as seen in figure 5.1), we can apply this to find a relationship between the CCM correlation and information as

$$I\left(X_t; \hat{x}_t\right) \geq -\frac{1}{2} \ln(1 - CCM_{X \to Y}) - D_{KL}(X_t||X_t^{\mathcal{N}}) \tag{5.16}$$

Another result from information theory (the data processing inequality (Cover and Thomas, 2006)) gives

$$I\left(X; f(Y)\right) \leq I(X;Y) \tag{5.17}$$

when $f(Y)$ is a function of Y that does not depend on X. More precisely

$$I\left(X; f(Y)\right) = I(X;Y) - I(X;Y|f(Y)) \tag{5.18}$$

is obtained from the mutual information chain rule and the fact that $I(X; f(Y)|Y) = 0$ (Cover and Thomas, 2006). The data processing inequality arises because mutual information cannot be negative. Hence, recalling that $\hat{x}_t = \hat{x}(Y_t^{(m)})$ is a function of $Y_t^{(m)}$,

$$I\left(X_t; \hat{x}_t\right) = I(X_t; Y_t^{(m)}) - I(X_t; Y_t^{(m)}|\hat{x}_t) \tag{5.19}$$

Combining 5.16 and 5.19 we find that CCM gives a lower bound on the cross-embedded mutual information CMI:

$$\begin{aligned} CMI_{X\to Y} &= I(X; Y_t^{(m)}) \\ &\geq -\frac{1}{2}\ln(1 - CCM_{X\to Y}) \\ &\quad -D_{KL}(X_t||X_t^{\mathcal{N}}) + I(X_t; Y_t^{(m)}|\hat{x}_t) \end{aligned} \tag{5.20}$$

Note that $D_{KL}(X_t||X_t^{\mathcal{N}})$ is a functional of the distribution of X_t only – that is, it is determined by the dynamical system which produces the marginal distribution of X_t, and does not generally change with the performance of the predictor $\hat{x}_t$ if the underlying dynamical system is kept constant.

To find conditions under which CMI is more precisely determined by CCM, we discuss conditions under which $D_{KL}(X_t||X_t^{\mathcal{N}})$ and $I(X_t; Y_t^{(m)}|\hat{x}_t)$ both reach zero.

For $D_{KL}(X_t||X_t^{\mathcal{N}}) = 0$, this occurs when X_t is approximately normally distributed.

For $I(X_t; Y_t^{(m)}|\hat{x}_t) = 0$, consider again an additive noise regression model, this time

$$X_t = \hat{x}_t + \eta \tag{5.21}$$

where η is an unknown noise variable. $I(X_t; Y_t^{(m)}|\hat{x}_t)$ is 0 if the noise η approaches zero, since

$$\begin{aligned} I(X_t; Y_t^{(m)}|\hat{x}_t) &= I(\hat{x}_t + \eta; Y_t^{(m)}|\hat{x}_t) \\ \lim_{\eta\to 0} I(\hat{x}_t + \eta; Y_t^{(m)}|\hat{x}_t) &= I(\hat{x}_t; Y_t^{(m)}|\hat{x}_t) \\ &= 0 \end{aligned} \tag{5.22}$$

This is a sufficient but not necessary condition to obtain $I(X_t; Y_t^{(m)}|\hat{x}_t) = 0$. It occurs when $\hat{x}_t$ is a very good predictor of X_t and thus η approaches zero everywhere. An alternative condition is that η is statistically independent of $Y_t^{(m)}$, so $I(\eta; Y_t^{(m)}) = 0$ and by the data processing inequality again $I(\eta; \hat{x}_t) = 0$ (since mutual information cannot be negative). Using this independence and the mutual information chain rule (Cover and Thomas, 2006)

$$\begin{aligned} I(X_t; Y_t^{(m)}|\hat{x}_t) &= I(\hat{x}_t + \eta; Y_t^{(m)}|\hat{x}_t) \\ &= I(\eta; Y_t^{(m)}|\hat{x}_t) \\ &= I(\eta; Y_t^{(m)}, \hat{x}_t) - I(\eta; \hat{x}_t) \\ &= 0 - 0 = 0 \end{aligned} \tag{5.23}$$

Satisfaction of this condition would be indicated by equal distribution of the residuals with respect to a linear function of $\hat{x}_t$ on a scatter plot like figure 5.1. In fact figure 5.1 only weakly supports this – there is a clear linear relationship, but the residual distribution appears to vary.

Substituting $I(X_t; Y_t^{(m)}|\hat{x}_t) = 0$ and $D_{KL}(X_t||X_t^{\mathcal{N}}) = 0$ into equation 5.20

$$CMI_{X\to Y} \geq -\tfrac{1}{2}\ln(1 - CCM_{X\to Y}) \tag{5.24}$$

Note that for the lower bound, we actually only require $I(X_t; Y_t^{(m)}|\hat{x}_t) = D_{KL}(X_t||X_t^{\mathcal{N}})$ to derive the above from equation 5.20. We have considered conditions under which both terms equal 0, however, if both conditions are violated, both terms increase since neither D_{KL} nor I can take negative values. Thus equation 5.24 may hold at least as an approximation under somewhat weaker conditions than strict adherence to those we have stated. This is likely important as it does not appear (from figure 5.1) that either condition can be expected to strictly hold, yet this bound does appear to apply in our numerical simulations in section 5.5.

5.4 Time delayed mutual information and transfer entropy

This section considers further bounds on CMI and briefly discusses the implications this has for both CMI and CCM as a useful measure for detecting causation.

Straightforwardly, CMI is bounded from below by the time delayed mutual information $TDMI_{X\to Y} = I(X_t; Y_{t+1})$, since

$$\begin{aligned} I(X_t; Y_t^{(m)}) &= I(X_t; Y_t, Y_{t+1}, \ldots, Y_{t+m-1}) \\ &\geq I(X_t; Y_{t+1}) = TDMI_{X\to Y} \end{aligned} \tag{5.25}$$

which derives from $I(A; B) \leq I(A; B, C)$ (Cover and Thomas, 2006). This is problematic, since TDMI is known to sometimes show non-causal or "super-luminal" information transfer (Schreiber, 1990). That is, $TDMI_{X\to Y}$ may be strictly positive even if X does not cause Y, due to temporal correlations introduced by an unmeasured "common cause" of the two. This has been cited as a motivation (Schreiber, 2000) for studying transfer entropy, which can be seen as TDMI with additional conditioning on the "past" information about the target variable (c.f. Granger causality as described in the introduction to this chapter or Granger (1969)).

$$TE_{X\to Y} = I(X_t; Y_{t+1}|Y^{(l)}_{t-l+1}) \tag{5.26}$$

In this formulation of TE a finite l-history is used for conditioning the target variable. We remark that if any common cause Z exists for X and Y, then $Y^{(l)}_t$ would include all information about Z_t, provided l is sufficiently large, by the same argument from the Takens delay embedding theorem that applied in our discussion of CCM. Moreover $Y^{(l)}_{t-l+1}$ contains this same information provided there is a strong auto-dependence in the time series Z and / or $Y^{(l)}$ (due to symmetry it does not matter which) over short (order of l) time scales. This, interestingly, relates the "Granger" sense of causation – statistical dependence of the cause on the effect, conditioned on all possible common causes (Holland, 1986; Barnett, 2009) – to the "dynamical" sense of causation described earlier and suggested by Sugihara et al. (2012). Specifically, we justify the use of $Y^{(l)}_{t-l+1}$ for "conditioning out" common causes for use in TE, by much the same logic that we justify the claim that $Y^{(m)}_t$ will contain information about the "true" cause for CCM.

5.5 Numerical comparison

This section attempts to validate our conclusions about CCM and CMI, and their relative performance as causality detectors relative to the more established transfer entropy approaches. Our hypotheses are first that CCM and CMI will generally behave similarly to each other, but with CMI bounded from below by CCM (section 5.3), and second that CCM/CMI will be subject to the spurious over-estimation of causal influences that we have noted are often cited as a motivation for the use of transfer entropy (section 5.4).

To test these claims, we compare CCM, CMI and two forms of transfer entropy on various coupled logistic map systems. In order to compare numerical results for CCM and CMI on similar scales, the plots for CCM are transformed by the formula $-\frac{1}{2}\ln(1-r^2)$, where r^2 is the original CCM value. Since all logarithms are to base e, we regard all mutual information / transfer entropy values as being measured in nats. The scaled version of CCM can also be regarded as implicitly measured in nats. Note that this scaled version of CCM is what appears on the right hand side of equation 5.24, and thus we know that under certain conditions it should be a lower bound for CMI.

Transfer entropy is calculated by two variants, required to deal with the fact that equation 5.26, being a continuous integral over probability density functions, cannot be directly evaluated from a data set. The first method is symbolic transfer entropy (Staniek and Lehnertz, 2008)

$$STE_{X\to Y} = \sum p(\vec{y}_{t+1}, \vec{y}_t, \vec{x}_t) \ln \frac{p(\vec{y}_{t+1}|\vec{y}_t, \vec{x}_t)}{p(\vec{y}_{t+1}|\vec{y}_t)} \tag{5.27}$$

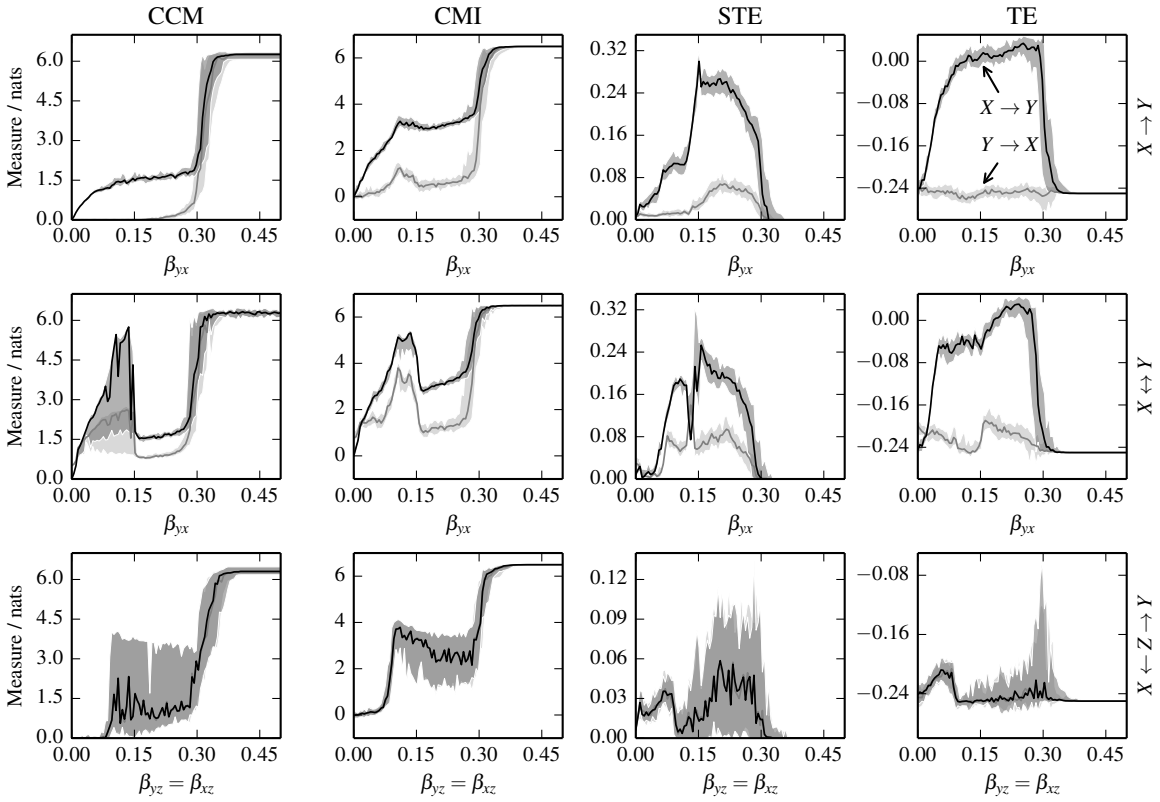


Figure 5.2: Causality measures applied to coupled logistic maps. Top row: unidirectional coupling $X \to Y$ in the dual system ($\beta_{xy} = 0$). Middle row: bidirectional but asymmetric coupling $X \leftrightarrow Y$ in the dual system ($\beta_{xy} = 0.02$, generally $\beta_{yx} > \beta_{xy}$). Bottom row: triple system with coupling from $Z \to X$ and $Z \to Y$ (but no direct coupling between X and Y). The four causality detectors are applied to each system with varying coupling factors and fixed $N = 1015$. CCM (left) is shown transformed by $-\frac{1}{2}\ln(1 - r^2)$, which scales it for better comparison to CMI. Black lines show median measures of causation from X to Y (e.g. $CCM_{X \to Y}$ etc), grey lines show measures from Y to X. Filled regions indicate 5th-95th percentile range from 20 runs of simulation.

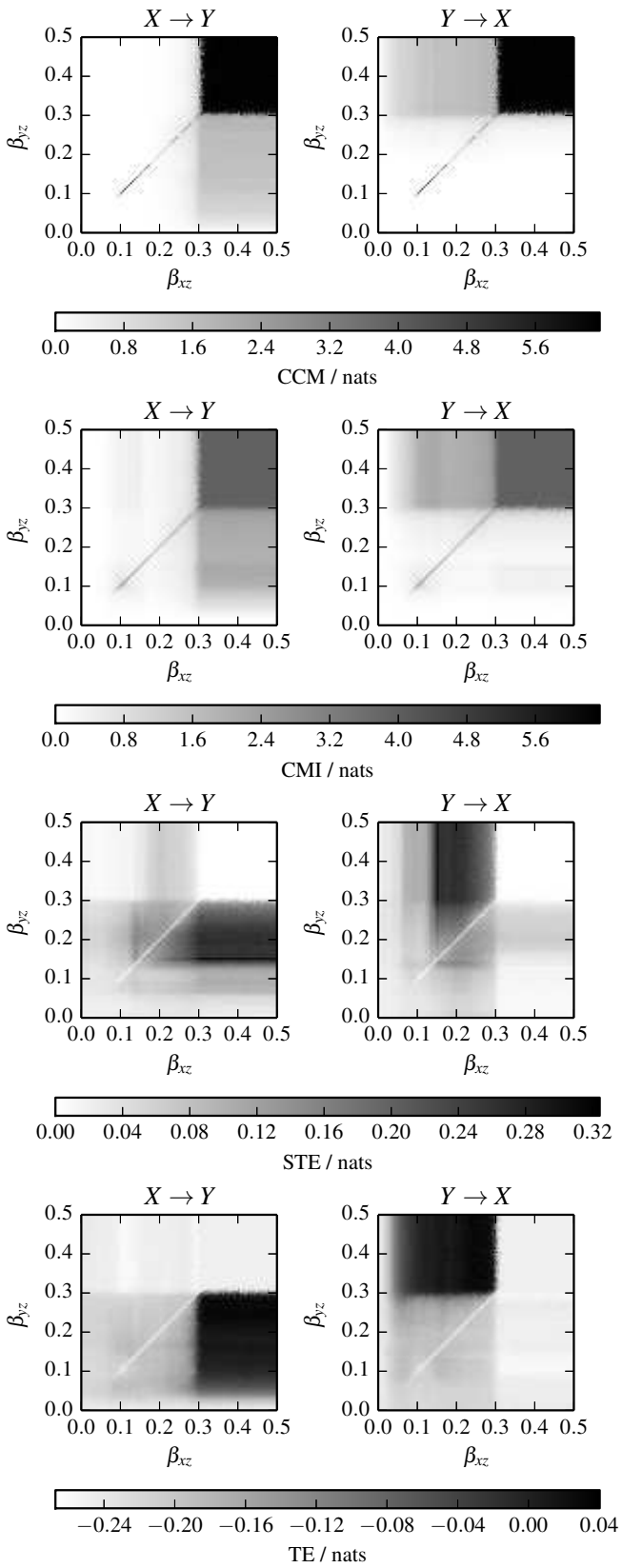


Figure 5.3: Causality measures in the trivariate system. Keeping fixed $N = 2000$, we vary both β_{xz} and β_{yz} in equation 5.28. Plots show median causality measures (a) CCM, (b) CMI, (c) STE, (d) TE from 20 runs of the simulation in each direction ($X \to Y$ and $Y \to X$). As before, CCM is transformed by $-\frac{1}{2}\ln(1 - r^2)$ and we treat the result as being measured in nats.

where $\vec{x}_t$ is the rank transform of vector $x_t^{(m)}$, i.e. the ith element of $\vec{x}_t$ is the number of elements which come before the ith element of the original vector $x_t^{(m)}$ according to a sensible ordering (e.g. ascending order, with a fixed rule for ensuring uniqueness in the case of multiple identical values in the original vector). Each rank vector now has a finite and discrete valued support, and so the information can be calculated using the sum in equation 5.27 (over all possible combinations).

The second form of TE uses k-NN estimation of conditional mutual information (Kraskov et al., 2004; Vlachos and Kugiumtzis, 2010) – which can be applied more directly to equation 5.26 – labelled simply TE in figure 5.2). The same k-NN mutual information estimator is also used for CMI.

We use $m = 4$ for CCM/CMI/STE and $l = 4$ for TE. The k-NN regression for CCM uses $k = m + 1 = 5$ (following Sugihara et al., 2012). The k-NN mutual information estimators (CMI/TE) use $k = 4$.

Some remarks about the bias apparent from the negative baseline values around -0.25 seen for this estimator in figure 5.2 (fourth column labelled TE) are warranted. Recall that transfer entropy is a conditional mutual information – the general form of which is $I(A; B|C)$. For example, suppose A, B and C are all mutually independently distributed random variables, then $I(A; B|C) = 0$, but also $I(A; B|B) = 0$, in the latter case, since we have conditioned on B, there is no way that we could have conditional mutual information even if A was statistically related to B. The k-nearest-neighbour based algorithm we use here for conditional mutual information has a known bias in the latter case ($I(A; B|B)$) of $-\frac{1}{k+1}$, but no bias for the former case where we took $I(A; B|C)$ with all variables independent. This means we cannot simply offset the known bias affecting the latter case without adding a bias which would affect the former case. It is to be expected that we would see this bias then for high values of coupling because we know that such dynamics lead to almost perfect correlation between the history of the target variable (used for conditioning) and the prior state of the causal variable – we have a case comparable to $I(A; B|B)$ where the conditioning variable effectively appears as one of the two variables. For moderate values of coupling, transfer entropy appears to "go up" more positive than this negative bias, but does not generally go above zero – it is possible therefore that we have a zero of the form $I(A; B|C)$ where we have three independent components in the conditional mutual information. We must therefore infer "positive" transfer entropy by a mixture of background knowledge. First, by comparison to the STE results which we expect to show similar behaviour but have no such biases. Second, our knowledge of the systems under study suggests it is unlikely that we would have the scenario where all variables are independent, since the conditioning variable is the past history of

one of the main variables, and there should be at least some statistical dependence between these two.

Figure 5.2 shows results from the dual coupled logistic maps from equation 5.4 as well as a triple system constructed analogously with three variables, X, Y and Z, where Z is a "common cause" of X and Y, that is, we have positive coupling $\beta_{xz} = \beta_{yz} > 0$ but there is no direct coupling between X and Y. The system equations are

$$\begin{aligned} x_{t+1} &= q x'_t(1 - x'_t) \\ y_{t+1} &= q y'_t(1 - y'_t) \\ z_{t+1} &= q z_t(1 - z_t) \\ x'_t &= (1 - \beta_{xz})x_t + \beta_{xz} z_t \\ y'_t &= (1 - \beta_{yz})y_t + \beta_{yz} z_t \end{aligned} \tag{5.28}$$

Note that we interpret the definition of causation (equation 5.3) as giving neither X a cause of Y nor Y a cause of X. For example, if we rewrite the RHS for x_{t+1} as a function $f_x(x_t, y_t, z_t)$, we find that changes in y_t cannot affect the value of f_x while holding x_t and z_t fixed, hence Y does not cause X. By symmetry, X does not cause Y.

The logistic map parameter q (in equations 5.4 and 5.28) is set to 3.7 to produce chaotic dynamics. Variables x_0, y_0 and where appropriate z_0 are initialised uniformly at random in $(0, 1)$, and the first 200 samples discarded to allow the system to reach its attractor before collecting data.

We have some empirical support for the lower bound in equation 5.24 since the plotted CCM and CMI values are generally similar, with CMI slightly higher than CCM. Note that the median values of CCM and CMI follow similar patterns, but there are clear areas where the CCM estimator has a much higher variance than the CMI estimator, suggesting that CMI is preferable.

These results point to a sensitivity/specificity trade-off between CCM/CMI and STE/TE as "causality detectors". In the unidirectional case (top row of figure 5.2), strong coupling ($\beta_{yx} > 0.3$)) leads to strong synchrony between the two systems. CCM/CMI interprets this as bi-directional causation, leading to a false positive inference of Y causing X, whereas STE/TE interprets this as all relationships being explained by a common cause, leading to a false negative failure to detect the true causal influence of X on Y.

The situation for CCM/CMI is worse than this though – in the three variable system under weak coupling (figure 5.2, bottom row, in the approximate range $0.01 < \beta_{yz} = \beta_{xz} < 0.3$), CCM and CMI report similar values as seen in the top row of fig 5.2 for the true causal relationship $X \to Y$ – but in the three-variable system there is never any direct causal influence between X and Y. Transfer entropy, on the other hand, slightly increases from the baseline for weak coupling in the three-variable system, but not to levels comparable to those seen when there is a genuine

causal influence.

For a broader picture of the results for the three variable system, we investigate combinations of distinct β_{xz}, β_{yx} for each of the measures. The median values of each measure are shown in figure 5.3. Recall that in this system, we have neither X causes Y nor Y causes X, and so any substantially positive values of the measures shown represent "false positives". The plots can be divided into four quadrants, with the lower left quadrant showing weak coupling for both variables (i.e. β_{xz} and β_{yz} both low), the upper right showing strong coupling for both variables and the remaining downward-diagonal quadrants show strong coupling to one variable and weak to the other.

One notable feature is the artefact that appears along the line $\beta_{xz} = \beta_{yz}$ in the lower left quadrant of all measures, suggesting all measures are sensitive to exact equality in the two coupling coefficients. However the artefact is inverted between CCM/CMI and STE/TE – i.e. exact equality in coupling leads to more false positives for CCM/CMI, but fewer for STE/TE. This effect was seen in figure 5.2 where $\beta_{xz} = \beta_{yz}$ was enforced.

In general, it is clear that false positive high values of the causality measures can occur in any of the measures. However, they are most likely when the coupling from Z to the causal variable (i.e. β_{xy} when measuring $X \to Y$) is high (seen in the right hand quadrants of the $X \to Y$ plots). This results from the fact that strong coupling from Z to X will lead x_t values to closely synchronise to the corresponding z_t values, meaning that measures of $X \to Y$ implicitly measure $Z \to Y$ (and there genuinely is a causal influence from Z to Y). However, CCM/CMI and STE/TE diverge when the coupling to the effect variable is also high (upper right quadrant) – by the same logic we are now implicitly measuring $Z \to Z$, which CCM/CMI gives high values to (due to the strong correlation) but STE/TE gives low values to (due to non-separability of Z from itself). In this case, what would otherwise appear to be a weakness of TE – underestimation due to non-separability – is in fact preventing TE from overestimating causal influence due to non-causal (super-luminal) information transfer between X and Y.

5.6 Conclusions

CCM and Granger methods appear to derive from very different assumptions about the nature of causation. However, adopting a "dynamical" definition of causation inspired by what appears to be the intent of the proposers of CCM (Sugihara et al., 2012), we can see that transfer entropy at least partially captures the same type of causal influence as does CCM. However the two methods are clearly distinct, and neither is perfect. This suggests that the distinction may be best thought of as a sensitivity/specificity trade-off. TE exhibits false negatives when common causes cannot

be easily separated from the direct causes – a well known issue with Granger methods (Granger, 1969; Sugihara et al., 2012). However, CCM simply trades this for the same problems as CMI and TDMI, namely misinterpretation of non-causal information transfer as causal influence, leading to false positives. This issues are most clearly present under strong synchrony, but in the contrary case of weak coupling we do not generally find a substantial difference between the results obtained from the various approaches.

It should be noted that in all cases false positives and negatives can arise for other reasons than those discussed. For example, common problems with transfer entropy include cross-talk in measurement noise (Vicente et al., 2011; Smirnov, 2013) and insufficient histories (Runge et al., 2012b; Lizier and Prokopenko, 2010), which can also lead to false positives. We suspect the cross-talk problem at least may also apply to CCM/CMI. We concentrate here on theoretical systems for clarity of argument – generalisation will naturally depend on application domain.

The analysis presented here hopefully helps to clarify what assumptions must be held to apply any of the tools discussed to detecting causation in the dynamical sense. Where CCM/CMI is to be used, we prefer our proposed information theoretic formulation, CMI, over the original definition of CCM from Sugihara et al. (2012). CMI can be described immediately in information theoretic terms (i.e. equation 5.1), and is also as a result straightforward to estimate using existing k-NN estimators for mutual information. The derivations in section 5.3 show that CCM could well be seen as a convoluted way of obtaining an inaccurate estimate of CMI, since both effectively measure the information that the (delay-embedded) effect variable contains about the causal variable (CCM finds this from the performance of an estimator trained using the effect variable, CMI finds it directly as the mutual information). We also note that in some cases we have studied (see figure 5.2) there is noticeably lower variability in the estimate of CMI as compared to CCM.

Chapter 6

Information dynamics of agents and hidden information

There is a close but non-trivial relationship between information transfer and causal connectivity, and recent research has used this to develop our understanding of the information dynamics of embodied agents. A known problem is that causal influences do not always show up as high information transfer even when they are present, but this chapter argues that if we complement a "complexity" oriented view of information with an understanding based on statistical inference, this becomes much less paradoxical.

This builds on the work in chapter 4. There we saw that physically strong causal influences may lead to low values of information transfer, in spite of the fact that information transfer has a meaningful justification as a tool for inferring the presence of causal influences.

From this perspective we can hypothesise the possibility of *hidden* causal information transfer. This occurs when causal links which are physically strong are not the points within a system where information transfer is most easily detected. For example, the causal influence of the state of the environment on a good, functioning, robotic sensor may be very reliable and almost deterministic, however, the causal influence of the state of the environment on a more downstream point (the robot's future actions, for example) may be less deterministic but more easily detected. At the end of this chapter, we see how this phenomenon can be manifested in a dynamical system. First, we will discuss how it is possible, and why it might, in particular, be relevant in the context of the dynamics of autonomous agents.

6.1 Introduction

A recent trend in studying agent behaviour has focused on the notion of information transfer in the sensorimotor loop (Lungarella and Sporns, 2006; Pfeifer and Bongard, 2007; Bertschinger et al., 2008; Williams and Beer, 2010a; Moioli et al., 2012). Furthermore, information transfer is sometimes thought of as relating (at least in a limited way) to causal connectivity. As a starting point, suppose that we accept that as much as sensor states *drive* motor states (as a control engineer might say), sensor states also (in some way) *cause* motor states.

It can be shown that various forms of mutual information can be used as an inference tool for causal relationships. Mutual information effectively measures statistical dependence between two variables and as such, if we accept the Reichenbachian notion that events do not predict each other unless they are causally related (e.g. one causes the other or there is a common cause), then high mutual information suggests some causal factor at work.

However, most recent research focuses on what I will call a "complexity oriented" view of information. From this perspective, we see that high mutual information generally corresponds to a more complex relationship between variables, and this is in turn assumed to suggest greater causal strength. In terms of agent behaviour, this might correspond to a relationship that appears "interesting" or even life-like.

While there is value in the latter view (complexity), adopting the former (inference) allows us to resolve certain confusions. The failure of standard measures such as transfer entropy (Schreiber, 2000) to capture the strength of causal influences has led some to develop more nuanced statistical measures with causality specifically in mind (Ay and Polani, 2008; Runge et al., 2012b,a; Sugihara et al., 2012) (see chapters 4 and 5). However, there is an important asymmetry that is made much clearer when we adopt the inferential view – there is no contradiction in regarding high information transfer as justifying the inference of a causal relationship, but allowing that low information transfer may correspond to either strong or weak causal coupling.

Importantly, causal influences can chain together – A causes B causes C and so on – or rather, the ability of A to cause Z might be mediated and facilitated by factors B, C, etc, that must come together in just the right way to make the overall causal influence possible. As shown later in this chapter, it is quite possible that the relationship between the end points can be "complex" while at the mid-points it is "simple" and thus is not detectable by looking at the information transfer alone. This leads to "hidden" causal connection, i.e. one which is real but does not "show up" as a high information transfer, even though the connection between the outer nodes does (Figure 6.1).

This chapter discusses the insights that can be gained from thinking of information both in terms of complexity and in terms of inference in section 6.2. Section 6.3 demonstrates hidden

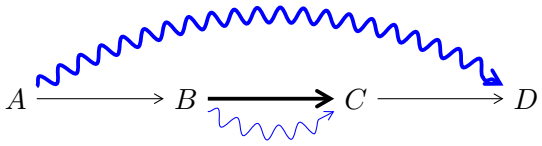


Figure 6.1: A conceptual graph of a hidden information. The thickness of the straight lines represents the physical strength of a causal coupling – strongest here between B and C. However since information transfer does not necessarily measure causal strength, it is quite possible that the information transfer (represented by blue wavy lines) is stronger from A to D than it is from B to C.

causation in a chaotic communication system, to elucidate the concept in a realistic but relatively simple model. Beyond this, chapters 8 and 10 investigate the possibility of hidden information transfer in more realistic models of autonomous agents.

6.2 Inference and Complexity

We have been using a *probabilistic* notion of causality, wherein X and Y are *prima facie* causally related (either one causes the other, or there is a common cause) if the two are statistically dependent (chapters 2 and 4). Importantly, since mutual information quantifies statistical dependence, it can be used as an inference tool for causal relationships.

First though, consider the "complexity" oriented view of information. Here, we view variables first and foremost as having their own *marginal Shannon entropy*, $H(X)$ and $H(Y)$. Variables that are more "uncertain" or disordered have higher entropy – a common interpretation being that the entropy measures the (theoretical minimum) number of *bits* (binary digits) that a description of the variable's value would require to communicate that value to an agent who had no prior information about the variable. Another similar mathematical result is that the Shannon entropy of X approximates the minimum size of a computer program that outputs a sequence of symbols drawn i.i.d. from the distribution of X, that is, the *Kolmogorov complexity* (Cover and Thomas, 2006, p. 473). The *conditional entropy* $H(X|Y)$ is the number of bits that is required to communicate a value of X to an agent which already knows the value of Y. The information is:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Thus there are two requirements for high mutual information:

- First term $H(X)$ or $H(Y)$: High entropy in the variables individually – one must initially be very uncertain about the value of X *or*, X must be "complex."

- Second term $H(X|Y)$ or $H(Y|X)$: Low conditional entropy – if one already knows the value of one variable, the value of the other is more or less certain *or*, the complexity of X is captured by Y (and *vice versa*).

The decomposition makes a very salient point. For the mutual information to be high, it is not enough that one variable can be predicted from another (captured by the low conditional entropy), there must also be a some variation or uncertainty in the variables to begin with (high marginal entropy). Furthermore, the information is bounded above by the marginal entropy or Kolmogorov complexity of a single variable (all entropies are positive and also $H(X|Y) \leq H(X)$, so $0 \leq I(X;Y) \leq H(X)$.)

Thus, when we use information to characterise agent behaviour, we often find that "complex" behaviours lead to high information. For example, one paradigm is the *predictive information* $I(S_t; S_{t+1})$ (Bialek et al., 2001) where S_t is a random variable representing the state of an agent's sensor at time t. Experiments have shown that where this is high, an agent will be simultaneously maximising both the temporal *predictability* of its sensor states as well as the *sensitivity* or *exploration* in its behaviour (such behaviour is called *homeokinesis* by Der et al., 2006, 2008; Ay et al., 2008). In other words, the agent's sensor state at time S_{t+1} should be at least somewhat predictable from its previous sensor state S_t to reduce $H(S_{t+1}|S_t)$, but it must avoid the trivial solutions (such as doing nothing) where the marginal entropy $H(S_{t+1})$ would be zero to begin with.

Predictive information is not usually thought to relate directly to causal influence, the most obvious reason being that the statistical dependence measured by $I(S_t; S_{t+1})$ is not enough to tell us that one causes the other – perhaps, for example, there is a common cause of the sensor states at both time points. In fact, this would be a natural explanation in the case of a passive sensor – a thermometer placed in a room for example simply reports readings S_t which are stochastic approximations to the true temperature of the room (call it the state of the environment E_t). Since fluctuations in temperature might be somewhat predictable over time, predictive information applied to the sensor $I(S_t; S_{t+1})$ might simply reflect the dependence in the environment: $I(E_t; E_{t+1})$. Moreover, if we condition on the state of the environment (somehow measured or known independently of the sensor state), we would expect $I(S_t; S_{t+1}|E_{t+1}) = 0$. This can be seen by applying the causal Markov condition (chapter 2) to the causal graph in figure 6.2a – the only path from S_t to S_{t+1} is blocked by E_{t+1}.

However, an active agent has its own changing state (e.g. the direction in which a robot is "looking", in combination with the state of the world, causally determines the information that

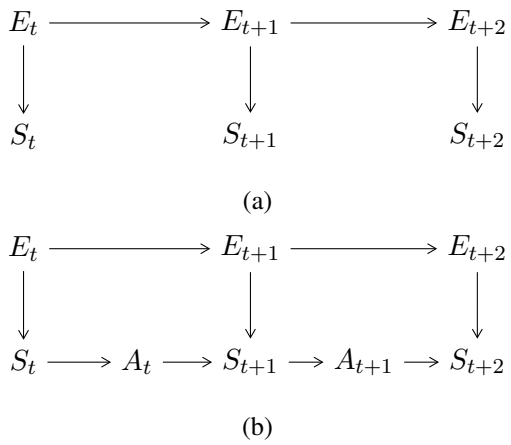


Figure 6.2: Causal explanations for predictive information. (a) Represents a static sensor where predictive information $I(S_t; S_{t+1})$ reflects temporal relationships in the state of the environment E_t. (b) Represents an active agent where sensor states are also related by a causal pathway through the state of the agent A_t.

appears at its visual sensors). Since the robot may react to its current sensor data S_t, the next sensor value S_{t+1} will be caused by a combination of the future environment state E_{t+1} an the "state of the agent" A_t which in turn is determined by S_t (as in figure 6.2b). The value $I(S_t; S_{t+1}|E_{t+1})$ may in this case be positive due to the causal pathway from S_t to S_{t+1} that is not blocked by E_{t+1}. Thus $I(S_t; S_{t+1}|E_{t+1})$ is a causal *information transfer* measure in the sense introduced in chapter 4 – that is, if S_t is not a cause of S_{t+1}, the true value of $I(S_t; S_{t+1}|E_t) = 0$, and thus a high measured value is grounds to infer S_t causes S_{t+1}.

The information transfer decomposes in a very similar way to the mutual information into two entropies, one representing marginal complexity and one representing dependence or prediction, except that now both are conditioned on the common cause:

$$I(S_t; S_{t+1}|E_{t+1}) = H(S_{t+1}|E_{t+1}) - H(S_{t+1}|S_t, E_{t+1})$$

Thus the "sensitive, but predictable" characterisation of predictive information also applies to information transfer in general, though we must bear in mind that if there are common causes, they must be considered. Notably, this means that if sensor states are strongly determined by the environmental state being measured, $H(S_{t+1}|E_{t+1})$ will be low, and therefore the information transfer will also be low, even if there is a genuine cause influence from S_t to S_{t+1}.

A more "interesting" information transfer would be $I(A_t; A_{t+1}|E_{t+1})$ – again an inferential statistic for the causal influence of A_t on A_{t+1}, and also a measure of the complexity of the agents state over time. From the complexity perspective, we can build a loose taxonomy of "low

information" behaviours based on the decomposition of information transfer to two entropies, illustrated in figure 6.3. That is, considering

$$I(A_t; A_{t+1}|E_{t+1}) = H(A_{t+1}|E_{t+1}) - H(A_{t+1}|A_t, E_{t+1}) \tag{6.1}$$

- Random exploration or "erratic" behaviour – the agent does not use the information from sensor states at time t to control the state that it will be in at time $t + 1$, it acts erratically. In this case, both $H(A_t|E_{t+1})$ and $H(A_{t+1}|A_t, E_{t+1})$ are high (near to each other), and the information is zero.

- "Inanimate" – the agent is simply not "doing" anything. In this condition internal state of the agent is not caused by its sensor states, and hence also is not caused by the state of the environment. In other words, $H(A_t)$ is always close to zero (for all t). Since $H(X|Y) \leq H(Z)$ (Cover and Thomas, 2006), both entropy terms also are close to zero.

- "Stable," or "doing nothing, for a reason" – the agent fixes its own state by responding in a very deterministic way to what is happening in the environment, not independently of it. That is, $H(A_{t+1}|E_{t+1})$ is low because the agent determines its state in a consistent way based on the state of the environment, not because (as in the inanimate case) $H(A_{t+1})$ is necessarily low.

The failure to distinguish between the latter two can be a problem – homoeostasis is an example of a useful "stable" configuration where an agent regulates some variable (e.g. body temperature) by reacting precisely to the environment: if the environment is too cold, the agent senses this and generates its own heat, to keep its body temperature (which in this case corresponds to its state A_t) constant.

On the other hand, the surface of Mars (at a given location over an appropriate time period) has a more or less constant "body temperature," but it is not self-regulated in the same way – if we could perturb the temperature of Mars somehow, there is no active regulation system that will counteract that perturbation – and our taxonomy would regard it as "inanimate." Note though that if we looked at the information transfer of the temperature from one time point to the next we would in either case find it was minimal.

Here we are not thinking of the agent's consciousness or other higher cognitive faculties (e.g. purposiveness or perception), and concentrate on the view that when the agent is "active" but "stable," there is still a *causal* connection between its sensors and motors. That is, at least in principle, if we were to intervene on the agent's sensors, its motor output might respond in some way (whether it would be conscious of this, or the agent would report that it did this *deliberately* or

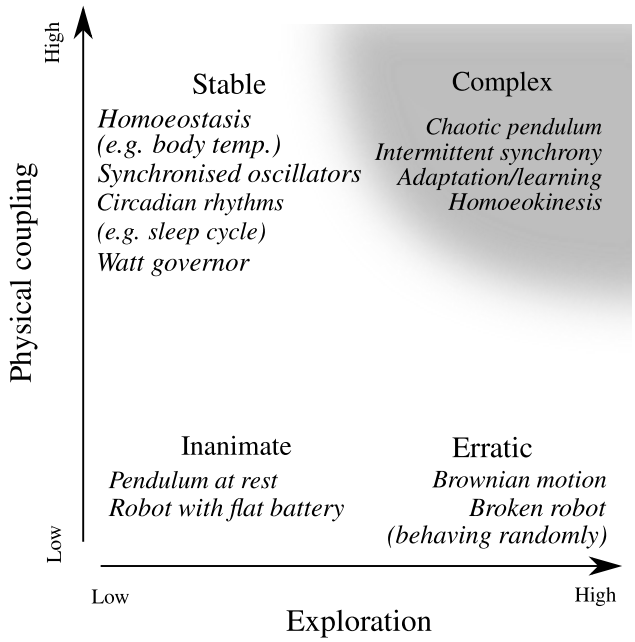


Figure 6.3: Taxonomy of behavioural complexity – an informal illustration showing how we might view different system behaviours in terms of information and causal coupling. The "exploration" axis is (loosely) related to the "marginal" term ($H(A_{t+1}|E_{t+1})$) in the information transfer equation – high exploration implies higher entropy. On the other axis, "physical coupling" loosely corresponds to the term that represents temporal dependence – $H(A_{t+1}|A_t, E_{t+1})$ – strong physical coupling would give a low value of this entropy. However, all information is positive, and thus the information transfer can only be high when exploration is high (equation 6.1). When exploration is low, the information transfer must also be low, even if (in a physical sense) coupling is strong (as in the "stable" region). We can only see high information transfer in the shaded "complex" region – where there is both high exploration and strong physical coupling.

not is not, for the moment, what we are interested in). Note also that the definition of "stable" here does not necessarily mean physically still (as we shall see, synchronous oscillation is an important type of stability) – only that there is not enough variation in the behaviour for causal links to be detected.

Returning to the view of information as a tool for statistical inference, the distinctions become more clear. In the case where the agent is stable, the information transfer will be low, but it is not that there is no causal connection between sensor states through time, it is only that, epistemologically, we do not have a way to *discover* those causal connections. In the case of complex behaviours, it is not necessarily the case that higher information indicates *stronger* causal influences, only that the causal influences that are present seem to be *easier to detect*.

To formalise the inferential view a little, we recap the model discussed in chapter 4, which took the "error statistical" model advocated by Mayo (1996) as inspiration. Suppose we wish to decide whether there is a causal coupling between the agent's sensors and motors. We might first propose the test procedure T which examines the hypothesis H:

- H: There is a causal coupling in the agent's sensorimotor loop.
- T: Observe the agent for a time and collect a data sample for the agents sensor values and necessary conditioning variables. Assuming stationarity and ergodicity[1] we can estimate from this an information transfer $\hat{I}(A_t; A_{t+1}|E_{t+1})$. If this is *higher* than some threshold I_0, infer that the agent is active, that its current state is a cause of its future state, i.e. "pass H," otherwise pass $\neg H$, that is, the alternative hypothesis that there is no causal coupling.

Suppose we perform T and find that it passes H. This alone is not enough, we need to also ask the question:

If H were false, what is the probability that T would still have passed H?

For the test to be a good test, the answer to this question must be a low probability – that is, the fact that the test indicated H is only good evidence for H to the extent that the test would not have indicated H if H were false (i.e. if $\neg H$ were true). This is what Mayo (1996) calls the *severity* requirement – if a hypothesis has passed a test, an inference towards that hypothesis is warranted as far as the hypothesis has been *severely* tested.

[1]These requirements relate to whether we can equate the proportion of time that a variable is found at some value in a time series with its probability of taking that value at any given time. These assumptions can be controversial though they are standard in information theory and time series analysis. In the cases we will study they are more or less guaranteed – for now, I will take them as acceptable.

In the case where the hypothesis is passed (i.e. $\hat{I} > I_0$), we can generally infer that a causal coupling is present, because we will generally only find $\hat{I} > I_0$ when the agent behaviour is complex (which in turn only occurs when there is a causal coupling). That happens because if there was no causal coupling (i.e. $\neg H$ was true), the true information should be zero: $\neg H$ entails $I = 0$, where I is the real information that $\hat{I}$ is an estimate of. Of course in a given data set, the estimated information may be greater than the "real" value as a result of "chance," but for a sufficiently high threshold I_0, the probability of this happening will be low. Therefore, if we find $\hat{I} > I_0$ then the hypothesis H that there is a causal coupling has been severely tested.

However, when the alternative hypothesis is passed, the question cannot be answered so easily. In order to establish the severity with which we pass $\neg H$, we need to know what the probability is of observing $\hat{I} \leq I_0$ under the assumption that H is true. But H (high causal coupling) could entail either stable or complex behaviour (refer back to figure 6.3), and we have no reason to expect one or the other. In the worst case, the alternative hypothesis could result in stable behaviour, in which case the expected predictive information would be zero. This would correspond to minimal severity – i.e. in a stable system where H is true, we would with high probability pass $\neg H$ using the given test procedure. Because of this low severity, we are forced to avoid drawing any inferential conclusions – that is we do not infer *either* H or $\neg H$, instead we accept that both are still possible – "no evidence of an effect is not evidence of no effect."

Considering some homoeostatically regulated system again – supposing we find it in a stable configuration where there is little variation in the variables of interest, the information transfer will be low, but this does not mean that there is not a causal connection. As far as we can tell, there is no way (from the existing experiment) to distinguish between stable, inanimate and erratic. However, erratic behaviour can be tested for simply by looking for random changes in the agent's state – the more problematic distinction is between stable and inanimate.

This inferential view of the information shows the behaviour of the agent in quite a different light. Considering only behavioural complexity we could only say that the causal coupling has an ambiguous relationship with information. But by combining complexity with an inferential perspective, we see an important asymmetry: high predictive information, i.e. behavioural complexity, corresponds to a state where causal coupling can easily and confidently be inferred – behavioural complexity makes the causal coupling *detectable*. On the other hand, low information tells us little about the causal couplings in the agent – it is quite possible either that there is low causal coupling, or on the other hand that the causal coupling is strong, but the behaviour of the agent is such that this fact is not obvious. In general, statistical dependence (mutual information) indicates some causal connection must be in operation, but the lack of dependence does not at all

indicate a lack of physical causation.

So a causal connection may be present even when it is not obvious from the data. But we are not stuck at this point. Just because a causal connection is not made immediately obvious by some data, does not mean that it can never be detected. Of course it may be possible to intervene on one of the variables and measure the response in another variable, but we will suppose for now that this is not possible and consider another case.

Imagine a causal pathway roughly represented by the chain $A \to B \to C \to D$. This could perhaps be a (very rudimentary) model of some information arriving at sensors at A, passing through intermediate nervous system components B and C, and affecting a motor response at D. Later we will see how it is possible in cases roughly corresponding to this form of system that a causal connection that cannot be easily observed between B and C facilitates in some way the overarching connection from A to D. That is, the sensory information "piggybacks" on some spike trains, voltages or other signals in the nervous system to affect a motor state, but the "internal" causal connections ($B \to C$) are not manifested in high information. One way to think of this is that A to D connection is "complex" whereas the B to C connection is (more or less) "stable." Thinking of information as measuring connectivity strength, there would be an apparent contradiction here – a chain is only as strong as its weakest link – but thinking of information as an inference statistic, this is not a problem since we do not take it that low information transfer from B to C implies that they are not connected.

A particularly important type of "stable" behaviour here will be synchronous oscillation. Remember to say that B causes C we would like to eliminate possible common causes as well as simply correlate the two. The usual technique to achieve this in time series analysis is to use Granger causality or transfer entropy, that is to consider the dependence between B at some time t and C at a short time later ($t + 1$) with the history of C (at $t, t - 1, t - 2 \dots$) conditioned out. The problem now is that in cases where B and C are synchronous oscillations, it is generally not possible to determine from the data whether one is causing the other (or there is mutual coupling), or whether there is a common cause of both, because after conditioning on the history of C, the value of B no longer contains additional predictive information. Thus mutually-synchronising oscillations are (conceptually) a kind of "stable" behaviour which are epistemologically indistinguishable from the "inanimate" behaviour where both oscillators simply reflect a common cause (looking only at observational data).

6.3 Hidden information transfer in communication

In this section we will look at the transfer entropy in a simple coupled dynamical system representing a communications channel, and show how a hidden causal influence can manifest. It is possible for physically coupled chaotic systems to synchronise to each other (Pecora and Carroll, 1990). This phenomenon can be used to build a "secret"[2] communications system. Cuomo and Oppenheim (1993) demonstrated the principle by building two analogue electronic implementations of the Lorenz system, where one system is designated the transmitter and the other the receiver. One voltage is taken from the transmitter and can be regarded as a communication signal. This voltage is then introduced into the receiver circuit such that one of the dynamical variables in the receiver is "driven" by its counterpart in the transmitter. The result of this is that the two circuits produce identical chaotic outputs.

Now if a small perturbation is added to the transmitter output, it will be present in the communication signal, though it will be "hidden" by the much larger chaotic oscillation. However, the receiver circuit will still synchronise to the transmitter's chaotic trajectory, and moreover, it will effectively clean the signal of the small perturbation. That is, the receiver's reconstruction of the transmitted chaos does not include the message. Thus the message can itself be reconstructed at the receiving end by subtracting the recovered chaos from the transmitted signal.

When we look at the time series of the two chaotic circuits they appear at first glance to be almost identical, even (up to a limit) when we are adding a perturbation to the first one. In that case when we find that knowing the state of the transmitter does not help us to predict the next state of the receiver if we already know the history of the receiver, since the history of the receiver is itself perfectly predictive of the state of the transmitter. This leads to a transfer entropy of zero.

However, when we look at the information transferred from the message generator to the reconstructed message, we should of course see a high information transfer when the communication system is working well. This is a prototypical example of hidden information transfer – the transmitted message passes through the synchronised chaotic circuits to influence the received message, but the circuits themselves do not show any information transfer. This can only be explained by remembering that the information transfer should not be thought of as a measure of "causal strength", but instead as representing how detectable a causal influence is.

We can perform a chaotic communication experiment quite easily in a computer using a logistic map as a chaotic system.[3] The transmission carrier c^t is a time series generated by a chaotic

[2] Though not in a cryptographic sense "secure."

[3] Examples and code for a simplified version of this experiment developed by the author are on-line at `http://www.jellymatter.com/2012/01/04/a-secret-message-from-another-dimension/`

map $f : \mathbb{R} \to \mathbb{R}$:

$$c^t_{n+1} \leftarrow f(c^t_n) = 3.9c^t_n(1 - c^t_n)$$

Then, let us assume that the "transmitted message" t is a Gaussian random variable, which will be scaled by a constant A to make it much smaller in overall amplitude to the carrier signal. This is then added to the carrier signal to produce a communicated signal s:

$$s_n \leftarrow c^t_n + t_n/A$$

The receiver consists of another chaotic map where the communicated signal is mixed with the map's previous iteration using a coupling factor $\gamma \in [0, 1]$, to get the reconstructed carrier: c^r

$$c^r_{n+1} \leftarrow f(\gamma s_n + (1 - \gamma)c^r_n)$$

Then, the reconstructed message r is obtained by subtracting the reconstructed chaos from the signal as it was originally received, then rescaling:

$$r_n \leftarrow A(s_n - c^r_n)$$

This system matches conceptually the chaotic communication channel as Cuomo and Oppenheim introduced it. The set of assignments that form the mathematical construction of the system above can be translated into a directed acyclic graph (DAG) by the following method:

- Assume that for each variable on the left hand side of an assignment (e.g. c^t_{n+1}) there is a random variable represented by the equivalent capital letter (e.g. C^t_{n+1}) of which the original variable is a realisation.
- Construct a "motif" for the time steps represented by n and $n + 1$ using the given assignments, i.e. if the assignment for a has b on the RHS (e.g. $a \leftarrow b$) then draw an arrow from B to A.
- Assume that the motif repeats over all time-steps $n \in \mathbb{Z}$.

Figure 6.4 results from applying this method to the current system.

To get an understanding of the behaviour of the system, the relationships between certain variables are plotted in figure 6.5 for low $\gamma = 0.2$ and high $\gamma = 0.8$ coupling. On the top row are the correlations between the transmitted and received signals. For low coupling, communication is not successful – there is not a clean correlation between R and T, but for $\gamma = 0.8$, the correlation is strong and communication is successful.

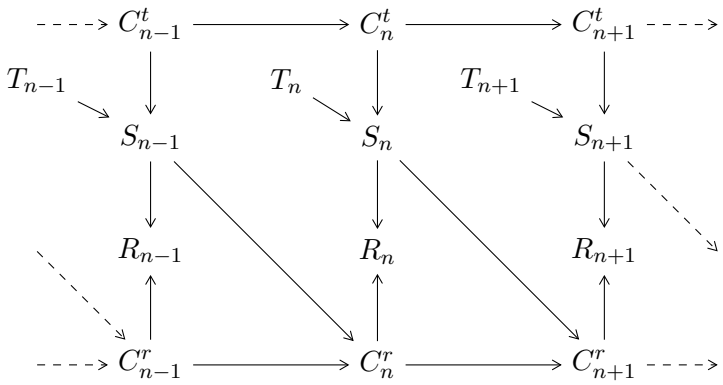


Figure 6.4: Causal DAG for the chaotic communication system. The transmitted signal T is introduced at each time step, and added to the transmission carrier C^t to get the transmitted signal S. The receiver synchronises its carrier C^r to C^t via S, which allows reconstruction of the message at R.

Similar relationships between the carrier signals can be seen in the bottom row of Figure 6.5 – for low coupling, though there is certainly a structured relationship between C^t and C^r, it is not strong enough to predict one from the other. But when the coupling is increased, and C^t and C^r are synchronised, the correlation between them is clear.

Thus it seems like our measure of "physical" coupling strength – γ – corresponds generally to the instantaneous statistical relationship. This is shown by figure 6.6 where the mutual information for the carrier $I(C^t_n; C^r_n)$ and message $I(T_n; R_n)$ is plotted – high values of the former suggest synchronisation, where high values of the latter show successful communication. Notice however, that though there is a broadly monotonic relationship between coupling strength and information, it is clearly non-linear – this shows the critical synchronisation point at around $\gamma = 0.4$. For lower strength couplings, there is neither synchronisation nor successful message transmission. Also, the message information is lower than the carrier information, reflecting the fact that although we can reconstruct the transmitted message at the receiver, there will be a lot of noise, whereas the relationship between the two carrier signals is almost perfect when there is strong coupling.

However, these information measures would not be said to measure causal relationships – just because two variables are correlated, we cannot say whether this is due to a common cause or a causal link in either direction between them. To look for causal relationships, we need to condition out common causes. We can do this by looking at the causal DAG in figure 6.4. Supposing we want to test the presence of the causal link from C^t_n to C^r_{n+1}, we can note that if this link was not present, the CMC would imply that $I(C^t_n; C^r_{n+1}|C^r_n)$ must be zero, which justifies it as a test statistic in the sense described above. Similarly $I(T_n; R_n|C^r_n)$ can be used to test for the influence of T_n on R_n. These statistics are plotted in figure 6.7.

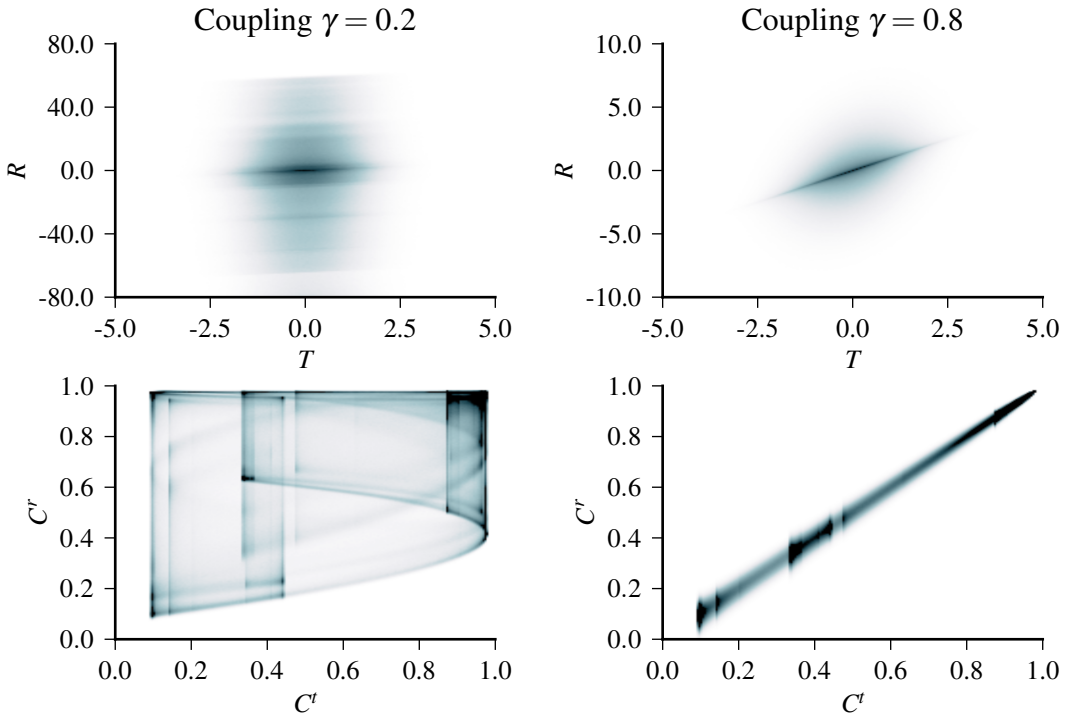


Figure 6.5: Correlations between signals in the chaotic communication system. The plots show two-dimensional distributions calculated empirically for the pairs (R, T) (message) and (C^r, C^t) calculated empirically from a large number of repeated runs of the chaotic communication system at two coupling values: low coupling, unsuccessful communication $\gamma = 0.2$ and high coupling, successful communication $\gamma = 0.8$.

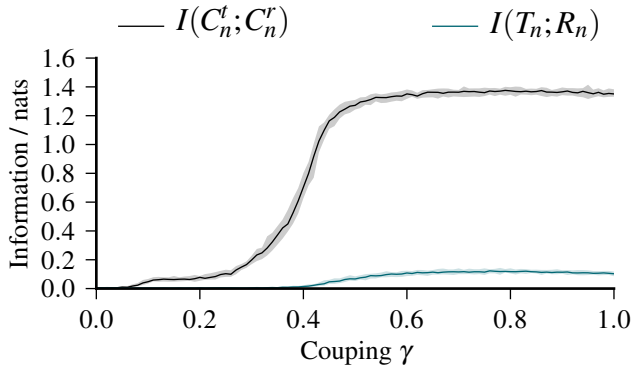


Figure 6.6: Mutual information in the chaotic communication system. This shows the statistical dependencies between pairs of variables – in black the carrier signals and in blue the message variables. Real values from the data are transformed to discrete values by constructing a set of bins in each dimension such that the distribution of data points is uniform in each dimension. For each value of γ, the mutual information is evaluated from a 30 time series of 2000 data points each with the starting values of the two chaotic maps chosen uniformly at random from $[0, 1)$. The central lines show the median value and the shaded regions show the 5-95th percentile range.

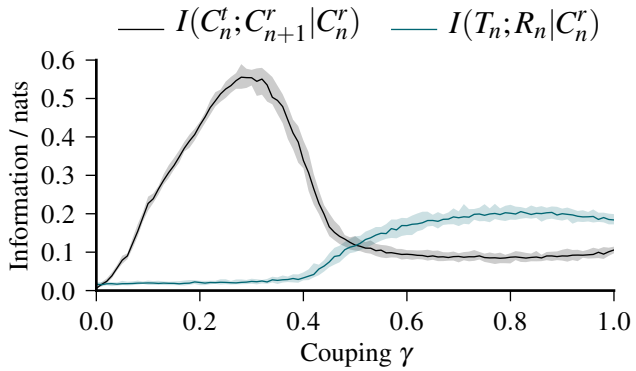


Figure 6.7: Information transfer in the chaotic communication system – from the transmission carrier to receiver carrier (black) and from the transmitted to received message (blue). Values are calculated and plotted using the binning and repeated sampling approach described in figure 6.6.

Notice now that the information transfer for the carrier signals is only particularly strong just below the synchronisation threshold at $\gamma = 0.4$. That is, is is easiest to detect the causal influence of the transmitted chaos on the reconstructed chaos when the synchronisation is incomplete (i.e. it is intermittent and "complex"), and in fact at values where the communication is not generally successful. When we go higher than the synchronisation threshold $\gamma > 0.4$ the information transfer from one carrier to the other actually decreases even though the coupling is stronger. On the other hand the information transfer for the message is higher than that of the carrier signal in this region, in spite of the fact that as we saw in figure 6.6, the carrier signals are much more closely synchronised.

The information passed from transmitter to receiver can thus be described as "hidden" at the point where its is transferred through the carrier signals. When the physical coupling γ increases to high values, the carriers are the most closely and directly coupled variables in the model – the information in the message must pass through the carrier signal to go from T to R. But the carrier signals show low information transfer when the message variables T and R show higher transfer – the information gets from T to R by "hiding" in C^t and C^r. This is intuitive when we think of the system as a secret communications channel – the message should not be immediately obvious from looking at the carrier signals. Moreover we see that the information transfer between the carrier variables does not monotonically increase with the coupling strength γ. However, the carrier signals are of course absolutely necessary for the message to be communicated at all – the causal influence of T on R, which is visible in the time series, is dependent on the causal influence of C^t on C^r, which is not (as) visible.

6.4 Conclusion

This chapter concludes the primarily theoretical part of this thesis. Its particular contribution has been to identify a possible phenomenon within information dynamics where information may be "hidden" at intermediate points in a causal chain. This follows directly on from the conclusions regarding the relationship between information and causal influences given in chapters 3, 4 and 5. However the proposed pattern of hidden information is a significant extension to the previously given theory – it is theoretically well grounded, but is also suggestive of empirical questions that will be studied in the remainder of this thesis. We have described hidden information transfer in the context of a communication system here since it hopefully offers a clear intuition of the information dynamics. The theoretical grounding also suggests the type of scenarios in which we would expect this phenomenon to occur – i.e. settings where we have "stable" behaviours characterised by strong, largely deterministic physical coupling. The example discussed above considers

synchronous oscillation as a facilitator of communication. In the second part of this thesis we will consider manifestations of the synchronisation phenomenon in systems geared towards generating stable gaits for robot locomotion.

Part II

Experiments

Chapter 7

Approaches to locomotion control

The previous chapters have looked in a largely theoretical context at the application of information theoretic analysis to discovering causal relationships in physical systems. In particular, this has been oriented towards robotic applications in general – dealing with questions of control and synchronisation, but mostly without specific reference to a particular task. The second part of this thesis will consider applications of these tools in more realistic scenarios, particularly with reference to legged locomotion.

The theory chapters have focussed heavily on synchronisation as a phenomenon of coupled oscillators (and discussed the associated pitfalls for causal inference techniques), which will be highly relevant in this second part of the thesis. The approach to locomotion taken here is inspired by considering a group of legs as a set of synchronising oscillators.

Chapters 9 and 10 deal with a full locomotion generation system for a hexapod robot designed in this manner. Before that, chapter 8 tackles a more simplified problem of a single oscillator synchronising to its environment. This chapter sets the context for this part of the work by evaluating major approaches to locomotion control found in the literature. This review is relatively brief and is intended to give an outline of the major issues to consider. More complete reviews can be found in Floreano et al. (2008) which focusses on evolutionary approaches and Kajita and Espiau (2008) which covers "classical" (non-evolutionary) techniques.

7.1 Legged locomotion based on passive dynamic walking

There has been renewed interest in the past decade or so in using the principles of dynamic locomotion and compliant actuation in robot control. In modern scientific terms, passive dynamic walking dates back to the work of McGeer (1990), who designed a biped capable of walking in a reasonably human-like manner down a slope without the need for any computer control or energy

supply other than its own gravitational potential.

In more recent work Harvey et al. (2004) (see also Vaughan et al., 2004, 2005, 2014) compare the approach of McGeer to that of recent robotic biped walkers such as Honda's ASIMO (Hirose and Ogawa, 2007), which relies on high impedance actuators and complex foot sensors. They contend that the engineering approach in such robot designs is constrained by a traditional industrial focus on precise control of position, rather than control of force. Relying on such control methods results in "unnatural" peculiarities where the robot cannot match human walking behaviour. For example ASIMO needs to have its knees constantly bent while walking, rather than locked straight during the stance phase as in human walking[1].

In contrast to this, passive dynamic walking is presented by Harvey et al. as an approach that will lead to more natural seeming walks, because the gait is more closely coupled to the physical dynamics of the machine. The passive dynamic walker effectively implements a simple model of walking through its mechanical construction, rather than using computer control to force a robot to match a pre-determined (e.g. motion captured) gait.

McGeer and Harvey et al.'s walkers consist of biped mechanisms, though the approach is also valid for quadruped walkers (Remy et al., 2009). An obvious advantage to approaching the latter is that the overall stability of the agent is a much simpler problem, due to the increased floor contact area with more legs. One argument might be that solving the stability problem with extra legs simply makes the task too easy and thus not interesting, but even in a quadruped there is still a valid question of achieving the desired "natural" motion that is energetically efficient and mimics the behaviour of animals.

7.1.1 Dynamic walking and dynamical systems

Harvey et al. (2004) consider dynamic walking to be an example of the broader category of the dynamical systems (DS) approach to robotics, which they put in contrast to "Good Old Fashioned AI" (GOFAI). By the dynamical systems approach they mean the exploitation the physical properties of the agent in order to produce the desired behaviour; in their words, the agent "naturally does the right thing". GOFAI techniques on the other hand involve computation of the correct outputs (motor activations) based on given sensor inputs and internal states using a general computing machine, comparable to a universal Turing machine (UTM). On a real instantiation of a dynamic walker (e.g. Collins et al., 2001) there is clearly no such computer involved, and even in a simulated dynamic walker as presented by Harvey et al. the computer provides the simulation, but the

[1] It should be noted that more recently robots using similar engineering approaches to ASIMO such as WABIAN-2R (Ogura et al., 2006) and HUBO II (Kim et al., 2008) have implemented straight-leg walking.

robot behaviour is still the result of its dynamics rather than a computational process. Thus in the DS approach to robotics, "computation" – as it is traditionally defined as operations performed on a UTM – is not the basis of the control system.

It is worth comparing this view with that of Stepney et al. (2005), who argue for what they call the "*computational stance*". Their perspective is that the dynamics of the universe perform computations, and that the word "computation" thus need not be limited to refer to those systems that fit the "Turing paradigm" of abstract mathematical operations on symbols. This implies two potential views of the DS approach, as *non-computational* (Harvey et al.) or as *non-classical computation* (Stepney et al.). It is clear however that in both cases the authors intend to be contrasting their methods with traditional UTM-based systems. There are ongoing debates around the contrast between these two perspectives, however they appear to be broadly compatible in the context of the current work, so where other authors such as Lewis et al. (2003) refer to "physical computation" it seems reasonable to take this as referring in many respects to the same concept as Harvey et al.'s "non-computational" dynamical systems approach, because they mean computation in the sense of Stepney et al.

Another subtle but important point with respect to the definition of the dynamical systems approach is the implicit assumption of a limitation to continuous time systems. Harvey et al. (2004) are not mathematically explicit about their definition of a dynamical system and thus leave this point slightly ambiguous: "there is a finite (though extremely large) number of physical variables that could be picked out as relevant to the workings of the machinery of the brain; and these variables *continuously interact* [emphasis added] with each other ... [Suppose we] write down an equation for each one, that states how its *rate of change* at any instant can be given by a formula related to the instantaneous values of itself and some or all of the other variables; then formally speaking we have a Dynamical System," (Harvey et al., 2004). Assuming the above implies continuous variables with respect to time then it contradicts the more standard formal definition of a dynamical system which permits discrete time functions (see e.g. Wiggins, 2003, p. 1). Of course in a physical robot, time *is* continuous, but there are interesting cases where agents are developed in environments with discrete time dynamics – these inevitably tend to be computer simulations, but it is still possible to use the dynamical systems approach of "naturally doing the right thing" in discrete time simulated environments as much as is it is in continuous time simulated environments. This is important because investigating a discrete time system may simplify the use of, for example, information theoretic measures which can later be generalised to continuous systems (e.g. Klyubin et al., 2005; Jung et al., 2011).

7.2 Alternative legged locomotion approaches

7.2.1 Zero moment point (ZMP) dynamic control

The Honda ASIMO introduced above is an example of a zero moment point (ZMP) based biped design. This technique is one of the most common biped walking approaches, and was first implemented in a robot in 1985 (Takanishi et al., 1985). However Vukobratović and Borovac (2004) make a slight over-statement when they say it is "the only procedure for biped gait synthesis," given that we have the already cited a drastically difference approach in the form of passive dynamic walkers. A broader review of this and other techniques for biped control is given by Kajita and Espiau (2008).

Briefly, the term "zero moment point" refers to the point on the foot-floor contact surface about which the torques in the floor plane are zero. In the limit of no acceleration in the robot body this becomes the projection of the robot's centre of mass to the floor plane. Provided the ZMP is well within the convex hull of the robot's foot contact areas on the floor the robot should not be inclined to tip on its feet.

Technically, the real zero moment point will never leave the support area created by the feet, because as the above implies, if it did the robot would passively tip over, and the pivot point (some point on the edge of the foot) would become the ZMP. Thus when the ZMP is measured using feet sensors, it is usually required that the ZMP have some distance from the edge of the foot to ensure the robot is balanced. In practise, robot gaits are generated through a mix of modelling the expected ZMP using Newtonian dynamics and correcting for errors with a feedback system, see Kajita and Espiau (2008) for a more complete overview of the implementation approaches.

It has been noted above that Harvey et al. (2004) point to several disadvantages of ZMP based control as compared to walkers based on the passive dynamic approach. Kajita and Espiau (2008) observe three properties of existing ZMP robots that could be seen as disadvantages: the need for six fully actuated joints (where passive walkers can use smaller numbers of under-actuated joints), use of position (rather than torque) control for the joints, and force sensors on the feet, arguably an unwanted expense.

Other drawbacks are the need for the feet to have a flat contact with the ground, to not slip, and generally to be inelastic. A formally ZMP-based robot could not run as one foot must always be in contact with the ground[2]. These cannot really be counted as advantages for the passive dynamic walking model given by McGeer (1990), which as the name implies does not permit running either, and also assumes infinite friction and inelastic foot-ground contacts, and gets round the problem

[2]In fact, the ASIMO now claims the ability to "run" (Hirose and Ogawa, 2007), though in practice its time in the air is barely noticeable, whereas animal gaits clearly can involve significant leaps.

of the foot shape by assuming point or arc feet.

7.2.2 MIT LegLab robots

Some of these limitations in both the ZMP and passive dynamic walking approaches are mitigated somewhat in the systems developed in the MIT LegLab under Marc Raibert in the 1980s. Many papers can be found in the relevant conferences and journals[3], but the work is reviewed in depth in a book by Raibert (1986). Here the focus was less on achieving independent balance and stability (though that was obviously a concern) and more on making use of legs as an energetically efficient way of achieving speed and balance with simple control. These are among the few well known examples of legged robots that use springs and compliant joints in running mechanisms.

The MIT legged robots were developed by starting with a monoped hopper in a 2D plane (Raibert and Brown, 1984). The dynamics for this robot are in fact quite straightforward and lead to a relatively simple control system for setting the hopping height, forward speed and torso attitude. It is shown that this control system can be extended to control biped and quadruped robots without too many additions – essentially an MIT biped is two monopeds attached at the hip and a quadruped is four (Raibert et al., 1986). The hopping robot was also extended to a 3D implementation capable of balancing itself in real implementations without extra support (Raibert et al., 1984).

The MIT systems are capable of much more natural running than ZMP based systems. This is partly a result of the design methodology since in a hopping robot the only way to move is to take the only foot off the ground for a part of the cycle. It was also found that it was easier to extend the control system to support quadruped locomotion when only one foot was on the ground at the time, as it introduces fewer new constraints (Raibert et al., 1986). Contrast this to the "crawl gait" that is typically considered the most stable quadruped gait (Kajita and Espiau, 2008) in which only one foot is allowed to be off the ground at any one time.

This is no accident. Raibert (1986, chapter 1) makes clear that the aim of the MIT research is to develop "active balance" in robots, which allows for faster movement. The ZMP approach is also a form of dynamic balance, because it takes into account the forces due to accelerations in the robot's gait as well as those due to gravity. However, it requires more constraints to be placed on the robot design and more complex feedback from sensors as described above.

For example, the LegLab's original hopper sets its height in a simple manner by applying constant force to the leg actuator on each cycle, and allowing the robot's physical springiness

[3] A publication list is available at `http://www.ai.mit.edu/projects/leglab/publications/publications-main.html`.

to determine the resultant height. Over time, the hopper will reach a height equilibrium for a given force. It would of course be in theory possible to set the height on each hop exactly (or nearly exactly) by measuring the forces on the robot and calculating an update based on a model of the robots dynamics, but simply relying on the spring to set the height is orders of magnitude simpler and does not require any complex sensors. This is a very straightforward example of the kind of physical intelligence that can be used in compliant robots, but this aspect of robot building is fundamental to the MIT approach. Later research investigated explicitly the notion of "self-stabilization" (Ringrose, 1997) – responding correctly to disturbances purely mechanically, without the need for feedback sensors. More recently this property has been investigated in human limb structures (Blickhan et al., 2007).

A particularly relevant example of physical computation was presented by Murphy and Raibert (1985) (see also Raibert, 1986, chapter 8). In a model of a trotting quadruped, it was shown that the stabilisation of the robot was dependent on the ratio of the hip spacing to the moment of inertia of the body. With the hips sufficiently far apart the robot passively stabilises to an alternating rhythm of the front and back legs. In dynamical systems terminology, they identified a bifurcation point in the system's dynamics where a stable limit cycle representing synchronised motion of the two legs appears. Murphy and Raibert do not apply this terminology, but nonetheless this is a precursor or perhaps early example of what Harvey et al. (2004) would call the "dynamical systems approach" to robotics.

Both ZMP and the MIT systems were initially developed in the early 1980s, at least with respect to applications in robotics. But today, ZMP systems seem to have a higher profile in terms of well publicised robotics research (e.g. the ASIMO). According to Kajita and Espiau (2008) the MIT LegLab's work "was not really followed up, but was surely the inspiration for many researches on cyclic systems." There are nonetheless a reasonable number of systems that clearly owe some considerable debt to these designs. The Rabbit (Sabourin and Bruneau, 2005; Morris et al., 2006) is a MIT style walking robot that reacts to external forces, an important extension of the idea of active balance, and the BigDog (Raibert et al., 2008) produced by Boston Dynamics (a company founded by Raibert) makes use of some of the design concepts such as semi-compliant joints in the limbs.

There has also been renewed interest in studying robots with compliant joints more generally (Scarfogliero et al., 2009), and particularly with respect to self-stabilising quadrupedal running (Iida et al., 2005). As well as permitting more natural "running" behaviour as in the MIT robots, it has been argued that compliant joints provide a better model of natural walking in humans (Geyer et al., 2006).

7.3 Energetic efficiency in robot walking

Pfeifer and Bongard (2007) discuss many of the above concepts (dynamic balance, self-stabilisation and physical intelligence) in the context of "embodiment", by which they mean an approach to cognitive science and robotics that considers the body and environmental situation of an agent as a key part of its intelligence. This approach motivates much of the present work, and is often argued to lead to greater energy efficiency (a theme that will come up again in section 7.5). Intuitively, it seems likely that exploiting the natural forces created by the body mechanics would of course provide better energy efficiency that high impedance actuators, but ideally we should quantify this difference.

Gabrielli and von Kármán (1950) introduced a metric for the efficiency of transport vehicles called specific efficiency, which is the ratio of power consumption to the weight and speed of the vehicle. Subsequently similar analyses have been applied to animal locomotion (Tucker, 1970) and legged robots (Gregorio et al., 1997). Gregorio et al.'s form of the measure is:

$$\varepsilon(v) = \frac{P(v)}{mgv} \tag{7.1}$$

The specific efficiency at a particular speed $\varepsilon(v)$ is the ratio of the power output of the robot at that speed $P(v)$ to the weight mg times the speed v. Because the specific efficiency takes into account the weight of the robot, it is a fairer measure of energy consumption and efficiency when comparing robots against each other. Gregorio et al. use this measure on their "ARL monoped" robot, which is similar in design to the MIT hopping robot, and has $\varepsilon \approx 1$ (varying slightly with speed). They compare this to other powered robots and find it to be better than its competitors, though human walking and McGeer's passive dynamic walker have specific efficiency closer to 0.03.

Much the same measure is applied by Collins et al. (2005) to the powered versions of the McGeer passive dynamic walker that they present. They refer to the same ratio as specific energy cost, by effectively multiplying the numerator and denominator of the right and side of equation 7.1 by the duration of the movement to get equation 7.2, where E is the energy consumed and d is the distance travelled.

$$c_t = \frac{E}{mgd} \tag{7.2}$$

In fact, they further distinguish between the "energetic cost" c_{et} where the energy is the total energy consumed by the system, and "mechanical cost" c_{mt} which only counts the work done by the transport mechanism to maintain the motion. The former measure is perhaps more general,

but the latter will compare only the mechanical efficiency (ignoring power lost to e.g. inefficient electrical power supplies or computational costs). Gregorio et al. (1997) are using the mechanical specific efficiency, because they refer to the power *output* by the system. The mechanical efficiency seems to be the only fair comparison between un-powered systems such as McGeer's and powered ones since the latter would inevitably achieve worse c_{et} due to said inefficiencies in input power conversions. Furthermore, for passive dynamic walkers, the total energy input to the system is simply the gravitational potential lost, i.e. $E = mg\Delta h$ where Δh is the height lost in the direction of the gravitational force. If θ is the angle between the gravitational force and the direction of travel in which d is measured, then $c_{et} = \frac{mg\Delta h}{mgd} = \sec\theta$, so the energetic cost is purely dependent on the angle of the test slope and not at all on the walking mechanism provided the robot achieves stable walking at all. The mechanical measure c_{mt} gives a value based just on the fraction of the total energy $mg\Delta h$ available that is used to create stabilizing forces by the robot.

Of course measuring the force as needed to find c_{mt} without disturbing the system is potentially quite hard. For this reason Remy et al. (2009) compare passive dynamic walkers by measuring the stable speed on a given slope. This is fair so long as the systems have similar mass.

Collins et al. report c_{mt} for their active dynamic walkers between 0.055 and 0.08[4]. By contrast they estimate the ASIMO c_{mt} at around 1.6. Scaled by weight, ASIMO's walking requires work done equivalent to around 20 active dynamic walkers or 50 humans (from Gregorio et al. (1997) human $c_{mt} \approx 0.03$ at maximum efficiency).

7.4 Central pattern generators (CPGs) as a control paradigm

A central pattern generator (CPG) is a portion of neural circuitry responsible for generating rhythmic signals used in locomotion. The presence of CPGs is currently a well accepted explanation for gait generation in many animals, and CPG inspired techniques have been used in a number of published robotic studies in control. A recent review by Ijspeert (2008) covers both biological and robotic aspects of the literature.

In a robotic or artificial agent scenario, CPG control can refer to a broad category of systems that rely on oscillatory pattern generators as a central component of the motion control. The earliest example of an artificial CPG controller comes from a simulation study by Taga and Yamaguchi (1991). The pattern generator here is a specific model of neural oscillations developed by Matsuoka (1985), which is a common model used by many subsequent studies, though it is not necessarily the only possible model.

A particular focus of interest in CPG studies is the notion that a single CPG system can gener-

[4]They also report an estimated c_{mt} for a third walker of "≥ 0.02".

ate multiple behaviours. A real model salamander (Ijspeert et al., 2007) showed how a single CPG may produce both walking and swimming behaviours in response to variation of a drive signal. Similarly dramatic changes in undulatory behaviour have been shown to take place as a result of environmental parameter variation in models of the nematode worm *C. elegans* (Boyle, 2009). This ties the CPG approach to locomotion back to the dynamical systems approach to robotics of Harvey et al. (2004), because the substantive changes in behaviour arising due to parameter variations can be viewed as the creation or destruction of limit cycles and fixed points in the phase space of the system, i.e. as bifurcations.

Another key notion in the dynamical systems view of locomotion is that of synchronisation. In animals CPGs are capable of producing rhythmic signals in isolation, but are thought to synchronise to proprioceptive feedback where it exists (Pearson, 1995). The idea of entrainment of the CPG to bodily feedback signals fits with the embodied robotics approach and has been studied on simulated (Pitti et al., 2009) and real (Buchli et al., 2006) systems.

A further and critical aspect of synchronisation is the potential for symmetry breaking among a set of oscillators. For example, a hexapod "tripod" gait – one of the simplest statically stable gaits – incorporates a spatio-temporally symmetric pattern of synchronisation between each leg (where each leg is in anti-phase with the leg opposite and, if it exists, the leg in front and/or behind). An influential paper by Collins and Stewart (1993) describes how spatio-temporal symmetries equivalent to a variety of animal gaits can be obtained from sets of coupled oscillators with a symmetric pattern of coupling between legs.

Such symmetric patterns of coupling between distributed CPG oscillators have been used to successfully generate animal and robot locomotion patterns (Beer et al., 1992; Golubitsky, 1998; Golubitsky et al., 1999; Cruse et al., 2002, 2007; Campos et al., 2010; Schilling et al., 2013).

Reflexes are also often used in robotic locomotion systems. Unlike CPGs, reflexes refer to systems that produce motor output only in response to particular sensory signals: for example the forced extension of a leg when it is detected that it has reached a certain physical state. The MIT LegLab control system (Raibert, 1986) could be viewed as primarily reflex based, though it does not make particular use of nervous system analogies. A number of systems have been built with hybrid controllers combining reflexes with CPG style gait generation (Lewis and Bekey, 2002; Fukuoka et al., 2003).

These approaches will be investigated further in chapter 9 where a combined CPG and reflex based system will be used to build a distributed gait generator which also incorporates physical feedback from the environment.

7.4.1 Coupled chaos control

A related approach is the use of coupled chaotic systems to produce locomotion patterns. This technique was introduced by Kuniyoshi and Suzuki (2004), who present it as a method of "*immediate* behaviour adaptation." Fundamentally the technique is closely comparable to CPG based control methods but the oscillators are chaotic rather than more straightforward limit cycle systems. It is proposed that the chaotic oscillators will explore a more diverse section of the phase space, but unlike a random exploration, chaotic systems are capable of synchronisation (Pecora and Carroll, 1990; Rosenblum et al., 1996), and hence a number of chaotic CPGs may still produce coordinated behaviour. Like CPG control, the system is model free in the sense that the oscillators do not explicitly model the environment or any other part system – instead behaviour emerges from the ongoing interaction between the coupled components of the system both inside and outside the agent.

As with CPG controllers, chaos based controllers are likely to produce a number of different outputs as a result of varying parameter configurations, including cyclic dynamics and more chaotic processes. Exploiting these dynamics makes it possible to use a chaotic control system as a search method for obtaining optimal behaviours. Shim and Husbands (2010) developed a simulated swimming robot controlled by coupled chaotic pattern generators, and by varying the chaoticity of the pattern generators in response to a goal function are able to make the robot tend towards effective behaviours. Thus this control method has been shown to be a viable approach to online, model-free value based learning.

The entrainment of chaotic control systems is often described in terms of synchronisation. Chaotic systems can be synchronised either completely or only in terms of phase (so the phase angle of two systems matches, but the amplitudes do not necessarily). It is argued by Pitti et al. (2009) that the latter case of phase synchrony (PS) can be usefully exploited in robot control. Using a planar biped walker similar to that of Taga and Yamaguchi (1991) they investigate the phase synchronisation between the body and neural components. They find that the neural oscillator is entrained to the body dynamics in cases where the performance (here measured by duration of successful walking gait) is higher.

These ideas show promise but as yet have not been implemented successfully in real robots. Recent work by Pitti et al. (2010) investigated the control of an isolated air muscle in this manner, but not yet a full robotic system. The theoretical foundations of this control method are perhaps not fully developed, perhaps as there are relatively few examples in the literature of this approach. In the above cited examples, the viewpoint varies slightly between one of focussing on synchronisation of the chaotic systems (Kuniyoshi and Pitti) and searching among multiple stable dynamical

configurations (Shim).

7.5 Analogue neuromorphic control

There have been many implementations of robot locomotion control using neural CPG inspired techniques. Experiments are conducted either fully in simulation broadly following the method of Taga and Yamaguchi (1991), or by using embedded microprocessors to generate the CPG signal in a real physically embodied system as per, for example, Ijspeert et al. (2007). In either case the CPG is instantiated by using a dynamical model (i.e. set of differential equations) and integrating numerically with the Euler (or some similar) method.

In the context of investigating scientific hypotheses as in the above two cited examples, numerical integration almost certainly provides an adequate model of CPG dynamics, at least as long as reasonable care is taken in the design of such systems. From an engineering point of view, however, CPG based control has to compete in terms of cost and energy efficiency with alternative methods.

Mead (1990) has argued that digital computing methods have inherent energy costs that prevent them from matching the performance of neural systems. This is estimated to amount to an unavoidable 10^7 times higher energy cost for a single operation in digital electronic logic compared to the same in biological neural systems. This is largely derived from the need to compose digital operations (`AND` and `OR`) out of large numbers of fundamental components (transistors), which each individually require energy to be charged or discharged. If, by contrast, we make use of electronic devices' physically defined characteristics as the fundamental operations (e.g. using capacitance to calculate integrals, or Kirchoff's laws to perform arithmetic), there is potentially a factor of 10^4 to be saved in energy costs. Mead advocates the use of analogue electronics to process information in a manner more directly comparable with biological nervous systems.

Mead bases his argument on the comparison of the energy cost for a single "operation", but this is perhaps slightly problematic because an "operation" could mean something different in neural and digital systems – it seems reasonable to accept that in some contexts a digital computer is more efficient than a brain (e.g. for well defined numerical calculations). Mead is trying to be fair by allowing each class of computing system to use the energy available in the manner most natural to it. However if we try to make a direct comparison in the current context of CPG simulation, it is clear that the Euler integration method would suffer from even worse inefficiencies than Mead discusses, because each step of the numerical integration would require (at least) hundreds of digital operations.

The bipedal robot of Lewis et al. (2003) is one of the few demonstrations of legged loco-

motion controlled by an analogue CPG control system. Artificial integrate-and-fire neurons are used to generate oscillating and spiking signals. The CPG output is fed to servomotors driving a hip joint. The knee is a freely moving passive hinge joint (with a hard stop preventing forward overextension). Thus the robot is under-actuated and the knee joint is used as part of the physical computation (Lewis et al. provide an analysis of the dynamics of the knee and shank highlighting its importance to achieving a working gait). The authors consider analogue electronic computation to be a form of physical computation with only a cosmetically different substrate (solid state electronics as opposed to Newtonian mechanics). In that sense they are effectively comparing the efficiency gains expected from Mead's arguments to those expected in dynamic walking (see section 7.3). The use of physical computation in this experiment is limited however, as the actuation is performed by servomotors as mentioned, which tend to have high impedance dynamics and require some further electronic transformation of the control signal.

Coupling and synchronisation are also an important component of Lewis et al.'s work. An angle sensor in the leg is fed back to the oscillator such that it is entrained to the movement of the leg. This constrains the oscillator dynamics which would otherwise drift indefinitely. The two CPGs (one for each leg) are also coupled to each other, enforcing the anti-phase oscillation of the two legs. This is a fundamental component of the gait, as of course the legs must move in synchrony (albeit with a half-period offset between them).

In a quadruped, the number of synchronous gaits is larger and can be described in terms of the symmetries of multiple oscillators (Collins and Stewart, 1993). A robot developed by Still et al. (2006) is capable of generating the various quadruped gaits using an analogue CPG controller. Four simple oscillator circuits are coupled in a chain, which is shown to be a sufficient system for generating symmetric oscillation patterns. The particular gait pattern is selected by modifying a small number of oscillator and coupling parameters. Unlike Lewis et al.'s system, the DC motors driving the legs are directly controlled by the oscillator output here, which removes control layers and simplifies the system. However the physical system here is less complex and does not appear to provide any interesting physical computation – there are no knees controlling the extension of the limbs, they simply have magnetic feet providing traction on the stance (backward motion) phase of each leg only.

Relatively few examples of analogue CPG control of robot walking beyond those cited have been published, though neuromorphic oscillator circuits continue to be researched for this and related purposes. For example, Vogelstein et al. (2008) present work aimed at animating the limbs of a paralysed quadrupedal animal with a neuromorphic chip.

Chapter 8

Hidden information transfer in an autonomous agent

This chapter describes a critical element of the information dynamics of embodied agents as described by this thesis, which can be thought of as hidden information transfer. This phenomenon has already been discussed in chapter 6 in a theoretical setting, here it is demonstrated in a minimal model of an autonomous agent. While it is well known that information transfer is generally low between closely synchronised systems, here we show how it is possible that such close synchronisation may serve to "carry" signals between physically separated endpoints. This creates seemingly paradoxical situations where transmitted information is not visible at some intermediate point in a network, yet can be seen later after further processing. We discuss how this relates to existing theories relating information transfer to agent behaviour, and the possible explanation by analogy to communication systems.

This chapter is based on the paper published in the Proceedings of the 2013 European Conference on Artificial Life (Thorniley and Husbands, 2013).

8.1 Introduction

The dynamics of embodied agent-environment systems are increasingly analysed using information theory (Lungarella and Sporns, 2006; Pfeifer et al., 2007b; Bertschinger et al., 2008; Klyubin et al., 2008; Pitti et al., 2009; Williams and Beer, 2010a; Moioli et al., 2012; Schmidt et al., 2012). This chapter adopts this approach and demonstrates a phenomenon that is consistent with the analogy to communications, but thus far seemingly overlooked in studies of information transfer in embodied agents. We describe "hidden" information transfer in a simulated robot: strongly physically coupled parts of the system carry information between separated endpoints, without such

information transfer being visible between the carrier components themselves.

Information transfer is often characterised using forms of *transfer entropy* (Schreiber, 2000), itself a non-linear generalisation of Granger causality (Barnett, 2009). Information transfer from X to Y is quantified by the relative improvement in statistical prediction of the future states of Y when the current state of X is known in addition to the already-known historical states of the target variable Y. A known issue, and potential source of confusion, is that information transfer is not a measure of the *physical strength* of coupling – a common example being synchronised systems, where very high coupling may mean that two time series are almost identical, leading to little prediction improvement and hence low transfer entropy in spite of strong physical coupling – a phenomenon already noted in this thesis (chapters 3, 4 and 6). This is sometimes regarded as a failure of transfer entropy to properly capture *causal* influences (Ay and Polani, 2008; Lizier and Prokopenko, 2010; Janzing et al., 2013). The agent model used in this chapter will exhibit this type of phenomenon, but in addition we will show that although strongly coupled components may exhibit low transfer entropy, they may still act as information conduits, hiding information transfer between more separate components.

Our model is a reactive robot designed to behave like a child swinging on a swing. As the feedback gain in the robot's controller increases, a self sustaining oscillation is created. The agent has a simple neural model acting as its brain, which is connected to the environment via its body. The state of the agent's neural system cannot (physically) influence the environment apart from by first affecting its body. However, we demonstrate that information transfer can take place from brain to environment *without* information transfer from brain to body. This shows how information transfer can be hidden within the agent, and revealed by its interaction with the environment.

This is the key result of this chapter – information can pass through a chain of coupled systems, e.g. A to B to C such that there is a high information transfer from A to C but *not* from A to B, even though physically there is no alternative route. In the discussion at the end of the chapter we will consider how similar effects occur in communication systems by way of analogy to our agent based model.

This chapter is organised as follows: the next section below describes the model swinging agent and its general dynamical features. The analysis in the following section shows the information hiding phenomenon by analysing the information transfer between each component of the system. The final section discusses this result and considers the implications for the study of embodied autonomous agents in terms of information theory.

Table 8.1: Variables and parameters

Symbol	Type/Value	Description
θ	Variable	Angle of pendulum from downward vertical
ω	Variable	Angular velocity of pendulum ($d\theta/dt$)
r	Variable	Current pendulum extension
v	Variable	Rate of pendulum extension (dr/dt)
u	Variable	Force control variable – force on bob due to effector
F_a	Intermediate	Force on bob due to acceleration
F_s	Intermediate	Force on bob due to spring
A	Independent variable (0-80)	Motor neuron output at saturation
g	9.81	Acceleration due to gravity
b	0.3	Pendulum damping coefficient
ρ	2	Motor neuron sensitivity
ϕ	20	Control parameter
k	100	Spring force constant
c	20	Spring damping ($= 2\sqrt{k}$ for critical damping)

8.2 Reactive swinging agent

The system studied here is a simplified model of a child swinging on a swing. The swing itself will be modelled as a rigid massless rod attached to a fixed pivot at one end with a mass (being the mass of the agent) at the other end. The agent's motor control consists in its ability to move the mass up and down (towards and away from the pivot). There are two general ways to approach the dynamical modelling of such a system. It is possible to use a "kicked pendulum" approach where a periodic forcing function is used to perturb the mass (e.g. Belyakov et al., 2009). However it has been found that even though a pendulum can be made to swing this way, the limit cycle produced is in fact unstable, and thus this is in a practical sense impossible to achieve in the real world, suggesting that a better approach is the "self-excited" oscillator (Pinsky and Zevin, 1999; Zevin and Filonenko, 2007). Here the agent creates a positive feedback loop by adjusting the distance from the mass to the pivot point (e.g. by raising and lowering the centre of mass of the agent relative to a fixed attachment point at the end of the rod). This create a stable limit cycle as well as a resting point (where the swing is pointing straight downwards and there are no vibrations to amplify). Thus if the swing is given an initial "push", the movement of the agent will sustain the oscillation, hence the system is described as self-excited. This approach treats the agent as

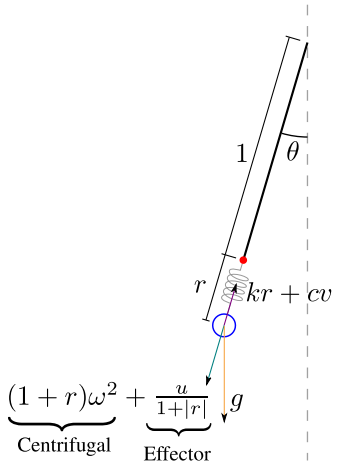


Figure 8.1: Spring based model of the swinging agent

a reactive system in the sense of Brooks (1986). This section provides further details on the implementation of this system.

A representation of the model is shown in figure 8.1. There is a massless rod with length normalised to one arbitrary unit. It makes an angle θ with the vertical axis along which the gravitational force g applies. The "agent" consists of a mass-spring-damper system attached to the end of the rod. The mass is influenced by the gravitational force, along with the centrifugal effect of rotation and the forces created by the spring: linear contraction kr where k is a constant and r is the extension of the spring, and damping cv with c another constant and $v = \dot{r}$ – the linear velocity of the mass in the direction of the spring. The agent creates an effector force u which acts on the mass, but this is reduced according to the current absolute extension of the spring, modelling a linear motor which produces less force output when it is already extended.

The full system can be described by the following equations. Table 8.1 lists each of the variables and parameters used. Dots represent differentiation with respect to a non-dimensionalised time variable t[1]:

$$\dot{\theta} = \omega \tag{8.1}$$

$$\dot{\omega} = -\frac{g}{1+r}\sin(\theta) - b\omega \tag{8.2}$$

$$\dot{r} = v \tag{8.3}$$

$$\dot{v} = \frac{u}{1+|r|} + F_a + F_s \tag{8.4}$$

$$\dot{u} = \phi(A\tanh(\rho v) - u) \tag{8.5}$$

The last equation describes the internal dynamics of the agent's reactive controller. The agent

[1]For simplicity all variables are treated as dimensionless, though the choice of $g = 9.81$ suggests the system could be treated as a one metre long pendulum with the agent mass at one kilogram, and time in seconds.

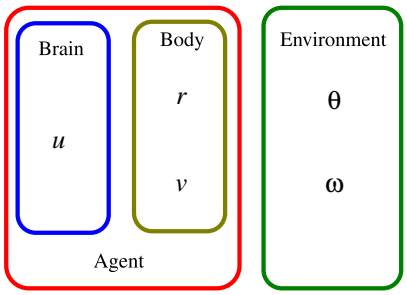


Figure 8.2: The agent and environment in terms of dynamical variables

senses the current velocity v of its spring, and passes this through a simple sigmoidal neuron, which determines a desired output force $A\tanh(\rho v)$ where A and ρ are parameters. The actual output force u moves towards this desired value in proportion to its current error according to the rate parameter ϕ.

The acceleration of the mass in the direction of the pendulum rod $\dot{v}$ is given by the resultant force (we assume the mass is normalised to one arbitrary unit). That is, equation 8.4 shows $\dot{v}$ is the sum of the force due to acceleration F_a (i.e. gravity and centrifugal forces, equation 8.6) and the force due to the spring F_s (equation 8.7) along with the effector force described above.

$$F_a = g\cos(\theta) + (1+r)\omega^2 \tag{8.6}$$

$$F_s = -kr - cv \tag{8.7}$$

We can treat the different dynamical variables as components of either the agent or environment, and further subdivide the agent into "brain" and "body" as shown in figure 8.2. The intention is to treat the agent as dynamical system which is "embodied" in the sense that its overall behaviour is a result of the close coupling of the agent's body, brain and environment (Pfeifer et al., 2007a). The main sensor variable is v – the input to the neuron, though the spring extension r can also be conceptualised as a component of the agent's sensory system. The motor output is represented by u, and the environment consists of the pendulum system: ω and θ.

The fixed parameter values used in the following simulations are shown in table 8.1. The parameter A effectively controls the feedback gain and will be varied as the independent variable in what follows.

The bifurcation plot in figure 8.3 gives an indication of the general dynamical features of the system. These plots are obtained by recording the angular speeds at which the swing passes through the downward direction, having been initialised with a random angular velocity and the "transient" time while the system is still far from a stable cycle or point discarded. Data is obtained

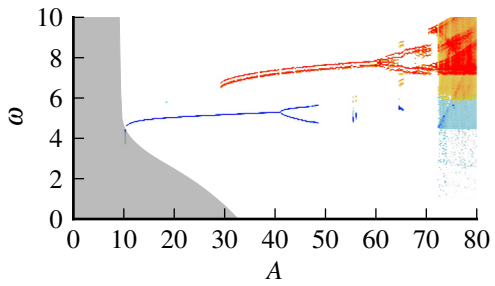


Figure 8.3: Bifurcation plot showing the behaviour of the system as the internal gain A is increased. Simulations are performed at each of 300 linearly spaced points between $A = 0$ and $A = 80$. Plot shows absolute angular velocity of the pendulum recorded as it passes through the "downward" ($\theta = 0$) plane of its state space – points in blue are returns to $\theta = 0$ where the sign of ω changed in between returns (i.e. the agent is swinging side to side) and points in red show returns in the same direction (the pendulum has swung over the top). The grey area shows the numerically estimated stability region for the fixed point where $\omega = 0$ – i.e. if the system is within this region it will eventually stop swinging. Outside of this region, it will go to either a swinging or rotating stable cycle.

using Runge-Kutta integration – all results in this chapter are based on an integration step size of 1/50th of a time unit, with a simulation length of 1000 time units. With A low, less than about 10, there is a single, globally stable fixed point – i.e. there is insufficient feedback for the agent to actually swing. Between feedback gains of around 10 and around 50, the agent usually swings side to side (represented in blue in the figure) – where the agent returns to $\theta = 0$ swinging in a different direction each time. Above $A = 30$ another stable cycle appears where the pendulum swings over the top rather than side-to-side, i.e. it returns to $\theta = 0$ travelling in the same direction (same sign of ω) each time. Note that the two cycles coexist between values of A around 30 to 50, but above that only the rotating motion occurs. Finally, above $A = 70$, a transition to chaotic motion occurs – above this point the system will sometimes rotate and sometimes swing side to side during a single trajectory. The fixed point where the system does not swing is locally stable for values of A less than around 33, meaning that sometimes the system will tend towards resting rather than either of the limit cycles. Thus the ultimate behaviour of the system is in general dependent on the initial conditions as well as the particular value of A chosen.

We now consider a slight alteration to the model. In practice, no sensor is perfect, and thus the input to the neuron might conceivably be modelled as a stochastic variable with a slight perturbation ε_v, so equation 8.5 becomes:

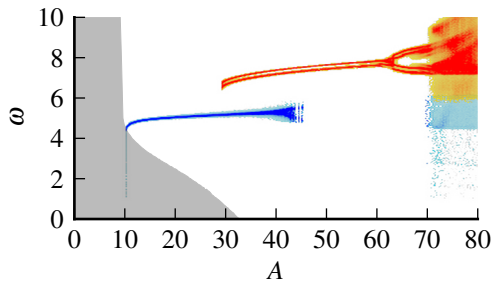


Figure 8.4: Bifurcations with noise $\sigma = 0.25$

$$\dot{u} = \phi(A\tanh(\rho(v + \varepsilon_v)) - u)$$

Assuming ε_v is small we can linearise its effect model it as a random additive perturbation on $\dot{u}$:

$$\dot{u} = \phi(A\tanh(\rho v) - u) + \phi A\rho\mathrm{sech}^2(\rho v)\varepsilon_v$$

In order to practically simulate the system, it must be written as stochastic differential equations. Specifically, we convert the equation for u (which is the only variable where we directly add noise) into *Langevin equation* form:

$$du = \phi(A\tanh(\rho v) - u)dt + \phi A\rho\mathrm{sech}^2(\rho v)\sigma dW$$

Where W represents a Wiener process, and a new parameter σ is introduced to control the strength of the random noise. This equation can be numerically solved using the stochastic strong order 1.0 Runge-Kutta algorithm. The full details of this approach including integration algorithm are found in Sauer (2012). The overall effect is that u behaves as if the neuron senses the current velocity v with additive Gaussian white noise, where the noise power is increased by increasing the newly introduced parameter σ. Figures 8.4 and 8.5 show the effect of increasing σ on the bifurcation structure – the main features remain much the same, but the crossing points are now somewhat random.

As well as making the model more "realistic", introducing this random perturbation ensures that the system is generally ergodic, which facilitates the correct calculation of transfer entropy. Without this property, the probabilities estimated from time series data tend to make little sense (see Breiman, 1969, for a discussion). This slight randomness also means that even for very closely synchronised variables there will likely be at least some transfer entropy measured, as there will be a constant introduction of entropy inside the system.

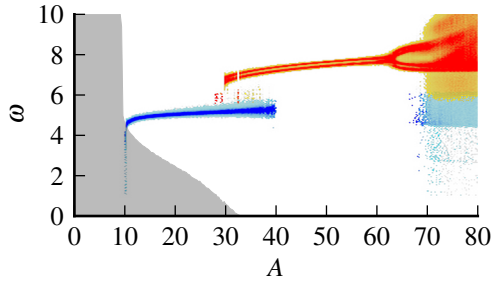


Figure 8.5: Bifurcations with noise $\sigma = 0.5$. Note that since the system is stochastic, the stability regions are not deterministically defined (close to the edge of the region shown, some trajectories may tend towards the fixed point and some towards the limit cycle depending on chance). The region shown shaded corresponds to the median stable boundary found in 20 simulation runs at each value of A.

8.3 Information transfer analysis

Transfer entropy is generally defined for two time series X and Y as a relative entropy or conditional mutual information:

$$TE_{X\to Y} = \sum P(x_t, y_{t+\delta}, y_t) \log \frac{P(y_{t+\delta}|x_t, y_t)}{P(y_{t+\delta}|y_t)}$$

The data points being taken at discrete time intervals δ, e.g. $X = (x_{t0}, x_{t0+\delta} \ldots x_{t0+n\delta})$. The sum is taken over the support of $P(x_t, y_{t+\delta}, y_t)$ – i.e. all possible combinations of values for the three variables. In this analysis we use the time interval $\delta = 1$ (i.e. 50 integration steps, corresponding to approximately one quarter of a cycle).

It is problematic to calculate the transfer entropy on continuous-valued time series such as we have here. We have used symbolic transfer entropy (Staniek and Lehnertz, 2008), which uses a convenient rank transform to find an estimate of the transfer entropy on continuous data without the need for kernel density estimation.[2] First an embedding dimension m is chosen (we use 4), for each $n \geq m$ we set $\hat{x}_{t0+n\delta} = \mathrm{rank}[(x_{t0+(n-m+1)\delta} \ldots x_{t0+n\delta})]$, where rank converts a sequence into its sort order, e.g. $(0.0, 0.4, 0.3, 0.25)$ becomes $(1, 4, 3, 2)$. That is, each original observation (after embedding in m dimensions) is a continuous vector ($x_t \in \Re^m$) and after transformation each observation is assigned one of the $m!$ possible permutations of a sequence of length m. The permutation is denoted $\hat{x}_t$ and for ease of calculation could obviously be assigned an integer representation according to an arbitrary one-to-one mapping. The formula for symbolic transfer entropy is then

[2] Alternatives exist such as k nearest-neighbour methods (Kraskov et al., 2004; Evans, 2008). At this time we are not aware of a reason to prefer one method over the other in this instance.

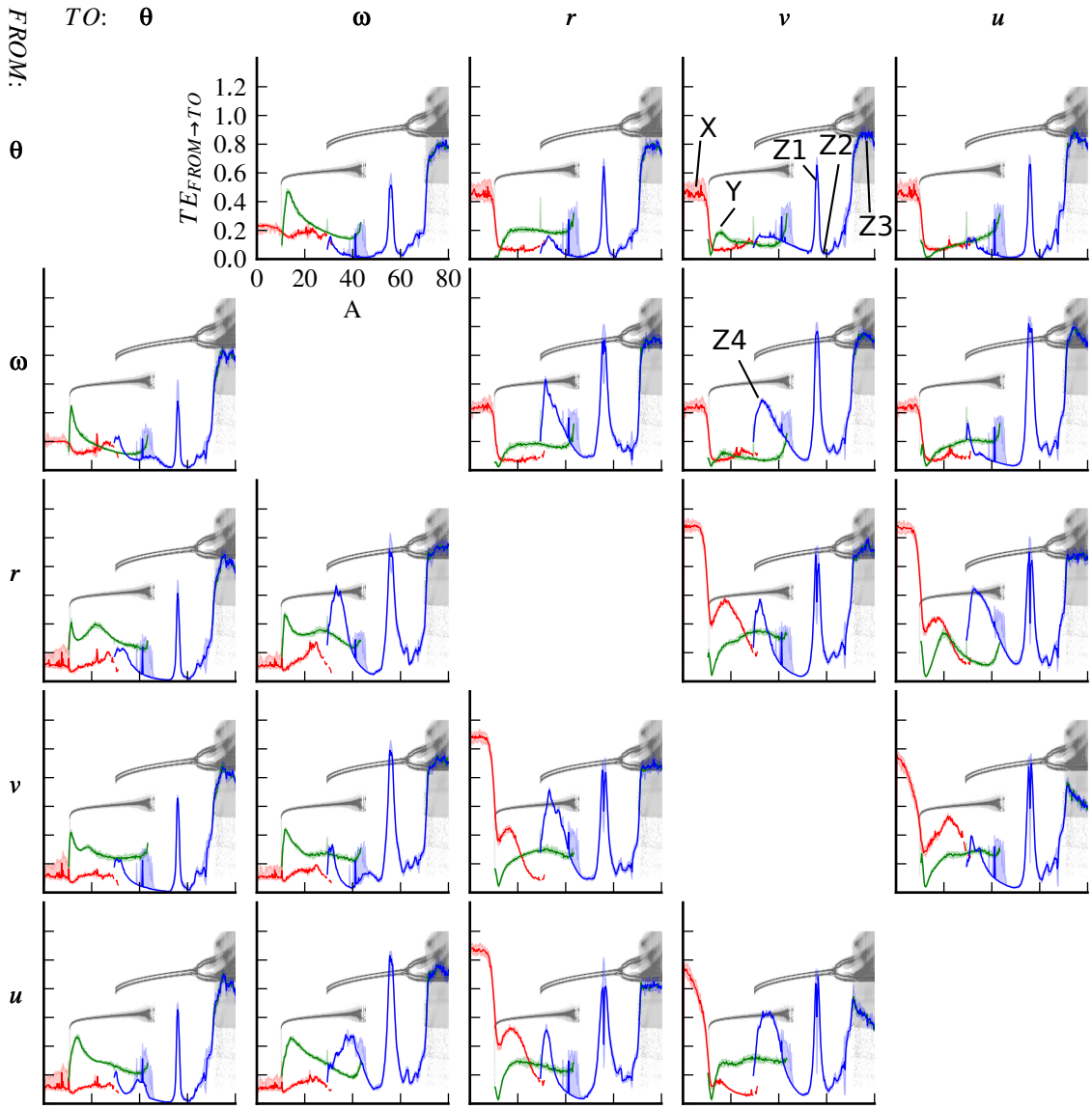
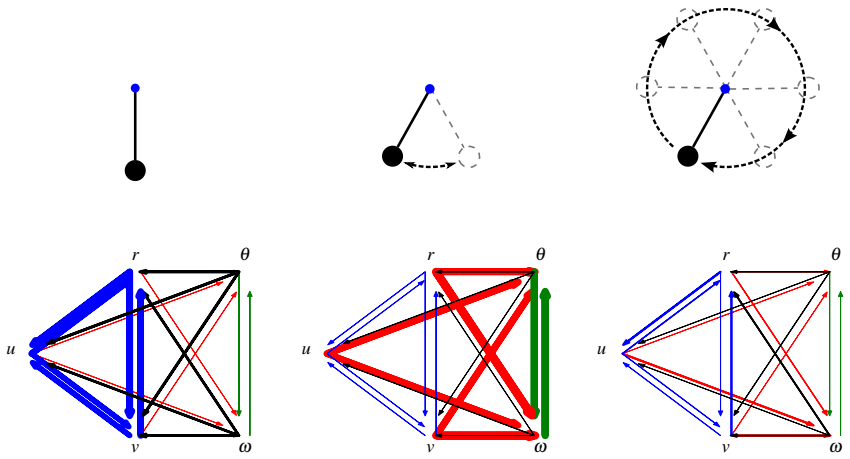


Figure 8.6: Symbolic transfer entropy in bits from each dynamical variable to each other one as the internal gain A is varied in the system with noise $\sigma = 0.25$. The background of each plot shows in grey a copy of the bifurcation diagram from Figure 8.4. The results from 20 runs are shown after grouping by behaviour mode: red for stable (non-swinging), green for side-to-side swinging and blue for rotational motion. For each behaviour the median is calculated for plotting and the shaded area around each line shows the 10th-90th percentile range where it is visible. The labels highlight the following features: at X the transfer entropy for low feedback gains (stable behaviour) is often high; at Y there is a peak in the curve for side-to-side swinging behaviour at around $A = 12$; Z1 and Z2 are a peak and a trough in transfer entropy that are not obviously related to the complexity of the behaviour seen in the bifurcation diagram; Z3 and Z4 show peaks in transfer entropy which seem to be related to features of the bifurcation diagram – the chaotic behaviour at Z3 and the bifurcation point at Z4.



(a) Stable (b) Swinging (c) Rotating

Figure 8.7: Median transfer entropy under three different behavioural regimes represented by arrow widths. Arrows are coloured blue for information transfer within the agent (variables u, r and v), green for within the environment (variables θ and ω), red for agent to environment and black for environment to agent. See Figure 8.2 for classification of variables. (a) Low feedback gain ($A = 2.6$), the system cannot maintain a periodic motion and tends towards the stable state. Higher transfer entropy is seen within the agent. (b) Moderate gain ($A = 12$), the agent will swing side to side. This graph illustrates the information hiding effect (see text). (c) High feedback ($A = 50$), the system rotates over the top. At this value almost no transfer entropy is seen in any direction. Note that the arrow widths in (a) are 1/3rd the scale of the widths in (b) and (c) since the transfer entropy values are generally much larger in (a).

$$STE_{X\to Y} = \sum P(\hat{x}_t, \hat{y}_{t+\delta}, \hat{y}_t) \log \frac{P(\hat{y}_{t+\delta}|\hat{x}_t, \hat{y}_t)}{P(\hat{y}_{t+\delta}|\hat{y}_t)}$$

With the probabilities estimated in the natural manner for discrete variables according to frequency of occurrence, i.e. $P(\hat{x}_t = X)$ would simply be the number of time points where $\hat{x}_t$ is found to be X divided by the total number of observations taken.

On every experimental run, the system is initiated with all dynamical variables set to zero except for ω which is taken uniformly at random from $[-10, 10)$. The first 100 time units are treated as transient non-stationary data and discarded, and the remaining 900 data points are fed to the symbolic transfer entropy calculation. This process is repeated ten times with different initial conditions, and the trajectories recorded are classified according to their final behaviour mode: resting, swinging or rotating.

The set of results in figure 8.6 shows all the transfer entropy values calculated for the system

using a noise amplitude of $\sigma = 0.25$, taking each possible combination of source and target variables. This shows a few basic features of the results. We see as expected that the transfer entropy does not straightforwardly correspond to physical coupling – there is no simple correspondence between the independent variable A and the transfer entropy value. We also see that very different patterns of information transfer are observed for the different behavioural regimes, even at the same value of A.

A simpler graphical representation of the transfer entropy is shown in figure 8.7. This shows the median transfer entropy for a particular behaviour at a chosen value of A as the width of an arrow pointing in the direction of information transfer. The arrows have been colour coded by the way in which they connect the brain, body and environment components.

The most striking result for our purposes is shown in figure 8.7b, where the feedback gain is moderate, resulting in a natural swinging behaviour. Here, the highest information transfer is along the paths coloured red which emanate from the agent (according to the classification in Figure 8.2) and flow towards the environment. This includes the arrows which directly connect the output of the motor neuron u to the environment variables θ and ω. However, there is no direct physical connection along this path since the coupling between the brain and environment is always mediated by the body. This is shown in equations 8.1 to 8.5 – the neuron output u does not appear on the right hand side of the equations for $\dot{\theta}$ and $\dot{\omega}$, and hence it can only influence these variables through the intermediate coupling to its body (since the body displacement r *does* influence ω). Thus the information transferred from u to ω (shown by a thick red arrow) for example is surely carried across the chain $u \to v \to r \to \omega$, yet there is *low* information transfer from u to v and r (illustrated by the thin blue arrows). It is in this sense that we claim this shows a form of hidden information transfer – we know that the brain can only influence the environment by going through the body, but even when a high information transfer is measured from brain to environment, there is a smaller amount from brain to body

Figures 8.7a and 8.7c do not clearly show this phenomenon, since it is in no sense necessary for it to be present. Figure 8.7a seems to show the strongest connections within the agent when the feedback gain is low and the system is resting, which can be explained by the fact that the source of entropy here is the sensor noise inside the agent, and since the agent is not swinging it may move up and down, but is not likely to influence the angle of the pendulum. In figure 8.7 the very high feedback coupling is likely creating a highly synchronised dynamic where the observed transfer entropy is very low.

8.4 Discussion

The key result of this work is shown in figure 8.7b, where during the entrained oscillatory motion of the system, the transfer entropy is shown to be higher from the brain to the environment than it is from the brain to the body, even though it is not possible for the brain to influence the environment without that influence passing through the body.

It appears that the entrained behaviour leads to a reduction in the transfer entropy measured within the agent, as can be seen by comparing the blue arrows between figures 8.7a and 8.7b. This is likely due to the close synchrony between these variables when the agent is swinging – a factor that is known to generally reduce measured transfer entropy. What is interesting is that though the swinging behaviour appears to decrease the transfer entropy within the agent, it also corresponds to increased information transfer from the agent to its environment. This is a clear demonstration of the importance of the agent's embodiment to the information dynamics of the system – the interesting (as in measurable) interaction takes place between the agent and the environment rather than within the agent.

What we are calling *information hiding* is the way in which information coming from a variable we specifically associate with the agent's neural system, i.e. u, appears to pass straight to the environment without having to "go through" the body, even though we already know that, in a physical sense, it must, since only the agent's body is physically coupled to the environment.

It is worth attempting to gain a little intuition for how this effect is working. For an analogy that is perhaps useful in the current context, consider the simplest type of encryption system based on a symmetric key illustrated in figure 8.8. A key is a randomly chosen binary sequence that has been previously shared between a sending and receiving party. The sender can encrypt a message by performing the XOR operation bit-wise between the key and the message. However, since the key was chosen randomly, the resulting encrypted signal should be statistically independent of the transmitted message – the encryption operation appears (to anyone without the key) to flip bits of the message at random (i.e. it randomly changes some 1's to 0's and some 0's to 1's). However, with the key, it is trivial to reconstruct the original message – the same XOR operation is simply applied using the previously shared key. Symbolically, if we have a transmitted message T, encrypted signal S and received signal R then we have a very low $I(T;S)$ yet high $I(T;R)$. This is a common feature of XOR type processes – for example, note that R is $S\ XOR\ K$ where K is the key, and so S allows one to calculate R if we already know H. Thus $I(T;R) = I(T;S|K) > I(T;S)$ – this increase from $I(T;S)$ to $I(T;S|K)$ (i.e. an increase in the mutual information when a conditioning variable is added) is an example of "synergetic" information (Williams and Beer, 2010b).

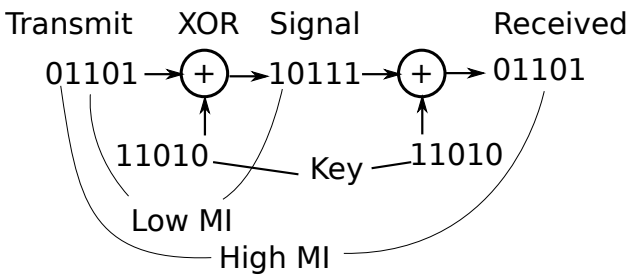


Figure 8.8: A simple encryption system

Though the above example was expressed in terms of mutual information rather than transfer entropy, this is closely related to the information hiding phenomenon as we have been discussing. Indeed if we assume the individual bits of the message and the key are independent of each other then $TE_{T\to S} = I(T; S|\mathrm{history}(S)) = I(T; S)$ and so on.

The information hiding process can thus be seen as a message being obscured by at some point and later reconstructed. In the example above this function is performed by the encryption system and is dependent on having a piece of secondary data (the key) shared between the two endpoints via some alternative channel to the main signal path. Of course, the encryption system is carefully designed to achieve this – it requires the deliberate sharing of the key and the use of the XOR function. However, comparable processes have been found relying only on chaotic synchronisation: the experiment of Cuomo and Oppenheim (1993) (discussed in chapter 6) demonstrated that synchrony between a pair of coupled Lorenz attractor systems can be used to "hide" information in a similar way. Their experiment, and our reconstruction of a similar experiment in chapter 6, suggests that it is plausible that information could be hidden by a dynamical process such as the one studied here without the need for the deliberate design of an encryption system.

We will consider this phenomenon again in a system involving a greater number of interacting components in chapter 10.

Chapter 9

Hexapod Locomotion

This chapter describes a reflex-based distributed controller capable of generating multiple hexapod gaits. The controller has been designed for application to a Trossen Robotics PhantomX hexapod. Gait generation is demonstrated here on the real robot and in a physically realistic simulation which has been tuned to closely mimic the properties of the real robot. The next chapter will examine the information dynamics of this system.

9.1 Distributed hexapod gait algorithms

The gait generator consists of a set of six independent relaxation oscillators built around each of the six legs, and a set of pulse-coupling influences which serve to synchronise the oscillators in a way that allows several different gaits to emerge, an approach substantially distinct from kinematic or central pattern generator (CPG) based architectures (see chapter 7). A number of similar systems have been published, the earliest directly comparable robotic system being that of Beer et al. (1992), though the current work is more immediately inspired by the more recent Walknet model (Cruse et al., 2002, 2007; Schilling et al., 2013), in particular in the internal coupling rules used to encourage emergent synchronisation between the distributed oscillators. However the current architecture is substantially simpler than the complete Walknet model and furthermore, unlike Walknet, is not intended to serve as a biological model. Nonetheless it does, like Walknet, aim to instantiate an *embodied* approach to robotics (Beer, 1995; Pfeifer and Iida, 2004) whereby the constant interaction of brain, body, and environmental dynamics is central to the generation of successful behaviour.

Before describing the architecture of the gait generating algorithm in more detail in section 9.3 the next section overviews the significant aspects of the robotic platform and associated physical simulation.

Figure 9.1: Trossen Robotics PhantomX Hexapod with electronics and batteries attached.

9.2 Robotic platform and simulation

The Trossen Robotics PhantomX hexapod (Figure 9.1) used here is an 18 degree of freedom hexapod of approximately 0.25m by 0.5m when the legs are laid flat. This section overviews the physical characteristics of the robot and the simulation used in this study.

A representation of the arrangement of the robot's legs relative to the main body is shown in Figure 9.2. The robot is, in effect, physically symmetric under reflection in both the saggital and transverse planes, and thus the choice of which end to regard as "anterior" is essentially arbitrary. Thus we pick one and assign a number to each leg as shown. For each leg, we define three joint angles for coxa, femur and tibia rotational joints, labelled $\alpha_i, \beta_i, \gamma_i$ where $i \in \{0, 1, \ldots, 5\}$ is the index of the associated leg. The orientation of these angles relative to the coordinate frame of a single leg is shown in Figure 9.3. The front and back legs are initially rotated slightly away from the centre of the robot by an offset angle $q = 17°$ (i.e. figure 9.2 illustrates the directions of the legs when all α_i are zero) – this mimics the physical construction of the robot.

In order to more easily collect large amounts of data and for prototyping the system, the robot was simulated using the Bullet Physics simulation engine. The robot was divided up into 10 ($1 + 6 \times 3$ for each leg) distinct rigid parts:

- The main body including coxa servos
- For each leg:

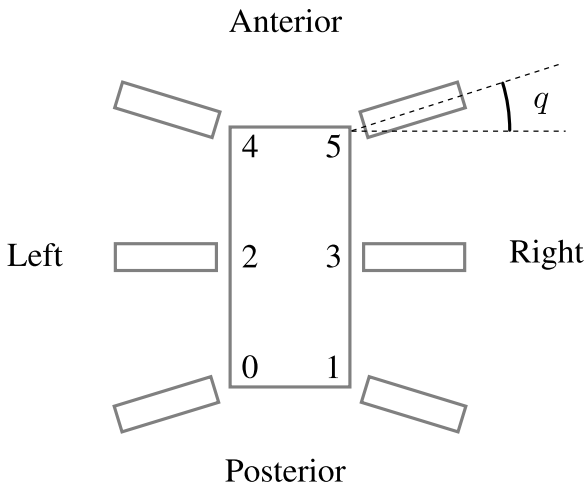


Figure 9.2: Top view of the hexapod. The numbering of each of the legs is used to index the joint angles (e.g. α_1 is the coxa angle of the right-back leg). The hexapod is treated as symmetric by reflection in the saggital and transverse planes, thus we treat "anterior" as meaning simply the normal forward movement direction of the robot. The angle q shown is an offset of the local proximal-distal axis for the front and back legs relative to the global left-right axis of the robot. That is, the coxa angles α_i (see figure 9.3) are taken relative to the orientations of the legs shown in this figure. This is in line with the physical construction of the PhantomX hexapod.

- The coxa-femur bracket
- The femur-tibia servo assembly
- The foot assembly

Each of these parts was disassembled from the real robot and weighed and measured along its major axes in the lab. In the physics simulation, the main body and femur-tibia servo assemblies are treated as individual cuboids, the coxa-femur bracket as a composite of cuboids, and the foot assembly is modelled using a set of three convex shapes (the same as those shown in figure 9.4). These shapes are used both for collision detection and for calculating inertia matrices. The centre of mass of each part is the central point of the cuboid for the simple shapes, and determined by symmetry and approximated by eye where necessary for the composite parts.

The physics engine provides a "debug" drawing output, a screenshot of which is shown in figure 9.4. This illustrates for clarity the complete collision model of the robot. For all experiments, the robot is placed on an infinite ground plane with gravity set to $9.81\mathrm{ms}^{-2}$ in the downward $(-y)$ direction.

To get an accurate simulation, a short internal time step of 5ms is used, and all dimensions are scaled by 20 (such that a "real world" metre is treated as 20 distance units in the simulation, applied forces and inertia matrices are similarly scaled as appropriate).

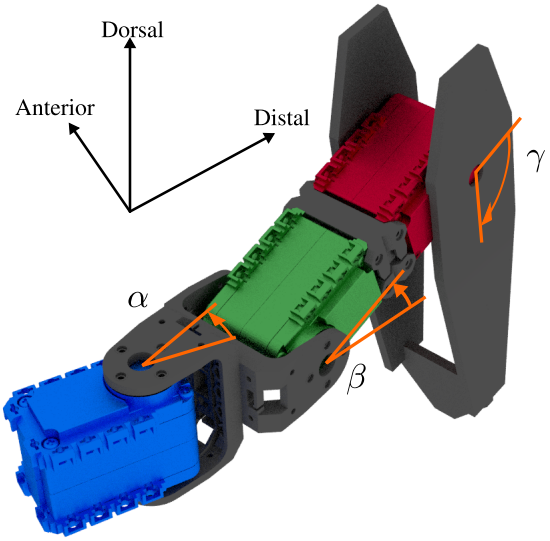


Figure 9.3: Illustration of a single leg showing the joint angles. The three servos that actuate the leg are shown (from proximal to distal, coxa in blue, femur in green, tibia in red). Each servo provides control over one of the three angles, α, β, γ (coxa, femur and tibia respectively). The direction of rotation for each angle is relative to the chirality of the anatomical coordinate frame shown. Thus, this image represents represents a right hand leg, whereas a left hand leg would appear as a mirror image with the "Distal" axis projecting to the left rather than the right. For example, in either right or left hand legs a small increase of α from zero will tend to move the foot towards the front of the robot – since of course a physical servo can only be rotated and not mirrored, these "modelling" angles do not necessarily correspond to the direction of positive movement as determined by the servo controller itself. An interface layer in our control system deals with any necessary angle transformations which are applied whenever angles are read from or written to the servo controller – for the purposes of modelling and the high-level control algorithm we treat all joint angles as behaving as shown. This image incorporates novel renderings of existing 3D servo and bracket models taken from I-Bioloid (2010, Creative Commons license CC-BY-NC-SA 3.0)

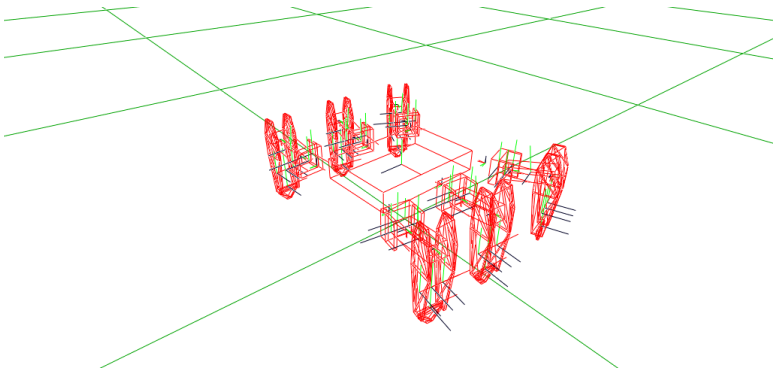


Figure 9.4: "Debug" output of physics simulation. Collision model is shown in red, along with ground plane and renderings of the local coordinate frames of each part. The shapes used in the collision model are also used for calculating inertial properties of the rigid body system in conjunction with the masses determined by lab measurements.

9.2.1 Semi-compliant joints in the real and simulated robots

All of the joints on the real robot are actuated by Dynamixel AX-12 servos (Robotis, 2006). These are used here in position-control mode, such that the output torque is adjusted relative to the position error $\varepsilon = \theta - u$, where θ is the current measured rotation of the servo horn, and u is the "set-point" – the desired shaft rotation of the servo horn, which is set by a control program and delivered to the servo through a serial interface. The angle θ is measured by a potentiometer built-in to the servo, and is also available to be read back to the main control program over the serial interface. The servo uses negative feedback to minimize this error, i.e. to set its actual output position as close as possible to the set-point.

Typically, position control mode is used for kinematic gait generation. That is, the desired position of the tip of each foot (relative to the body) at each time step is pre-determined for a given gait, and the corresponding positions for each servo (i.e. the configuration of servo positions that results in the desired foot position, according to the geometry of the leg) are calculated. These servo positions are sent from the controller as set-points to each of the servos, and the feedback mechanism inside the servo deals with ensuring that the servo reaches the desired set-point in a short time.

The gait generation approach we take is substantially distinct to this traditional model. We describe an algorithm which takes advantage of the *embodiment* of the robot, that is, where the interaction of the robot's "brain" (by which we mean control program) and body is key to the successful generation of the gait. This means that information about the current state of the body has to be available and relevant to the control program. In the traditional kinematic model, the control program is essentially open-loop – set points are generated by the control program and fed

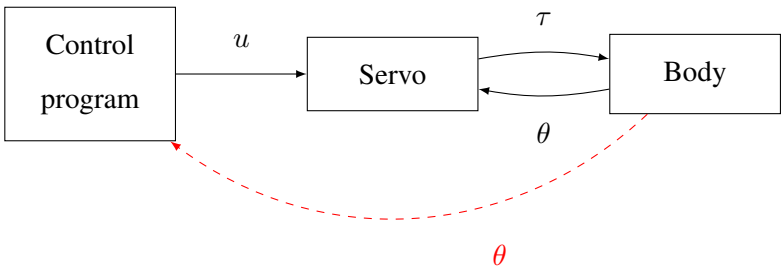


Figure 9.5: Incorporating bodily feedback in the control system. This diagram conceptually represents the distinction between an open-loop kinematic controller and the "embodied" approach taken here. The key difference is the feedback pathway shown in red dashes, representing information about the current state of the body being fed back to the controller. In a typical kinematic control system, desired joint angles u are given to the servo motor by the control program. A module within the servo then calculates output torques τ such that the true angle of the output shaft θ is as close to u as possible – typically negative feedback is employed here and thus the servo can be considered a closed-loop controller. However, the joint angle θ is *not* typically fed back to the main control program – i.e. in kinematic control the pathway shown in red dashes is not used. In contrast to this, the gait generator presented here makes use of this extra feedback pathway from the body to the main control program. Closing the loop between the body and the controller in this way is a key part of the embodied approach to robotics.

to the servo controllers, but sensory information about the current state of the body is not used by the controller. Our approach closes the loop between the controller and the body as shown in figure 9.5. The main control program can still only send set-points u to the servo controllers, but in addition reads back the current measured state of the servos (i.e. the measured output angle θ) and uses this information as part of generating subsequent control outputs.

However, as discussed, the typical position control system employed by the servos is designed to minimise the position error in as short a time as possible. In other words, under normal operation, the feedback feedback value θ read by the control program will be almost identical to the previous control signal u that the control program just sent, and so the feedback cannot contain any useful information for the control program.

To avoid the feedback information matching almost exactly the set points, there must be a non-trivial amount of *compliance* – meaning that the output shaft can deviate from the set point by significant amounts (in normal conditions) without being immediately corrected by the servo controller.

The Dynamixel AX-12 controller provides three settings for adjusting compliance (Robotis,

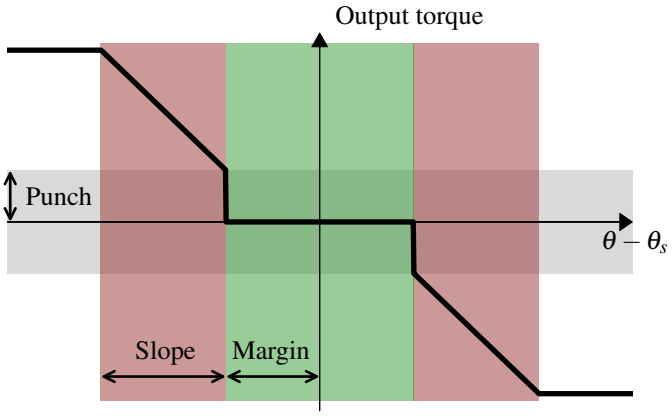


Figure 9.6: Compliance curve for AX-12 servo.

2006). The parameters *punch*, *margin* and *slope* adjust a compliance curve illustrated in figure 9.6. The specification for the servo (Robotis, 2006) is not specific about how this curve is used to control the output torque, however it appears that it determines the duty cycle for the input to the armature, thus indirectly controlling the torque generated by the armature. This naturally does not take into account frictional and other electromechanical forces that would affect (and generally reduce) the torque on the output shaft.

We set both punch and margin to zero, resulting effectively in proportional negative feedback, with a cut-off at maximum output power, and a slope set by the slope parameter. This parameter is allowed to take integer values in the range 0-255, with larger values corresponding to gentler slopes. To give an idea of the effect of varying this parameter, the step response of a femur servo (in the fully constructed robot) was measured on a single leg with coxa and tibia servos fixed at $\alpha = \gamma = 0$. For this purpose we regard the femur joint angle β as equivalent to the measured angle of the servo under test which we label θ. We set the control signal

$$u(t) = \frac{\pi}{2} H(t) = \begin{cases} 0 & t < 0 \\ \frac{\pi}{2} & \text{otherwise} \end{cases}$$

Where H is the Heaviside step function, thus the control signal tells the servo to move a quarter of a turn. In this experiment, the robot is raised off the ground by placing a rest underneath the main body such that the legs are free to move without interacting with the ground. The responses of the servo at several settings of the slope parameter are shown in figure 9.7. Decreasing the slope parameter value increases the strength of the negative feedback. Damping is introduced by fixed mechanical properties of the system (e.g. friction in the gear system), thus for a very low slope

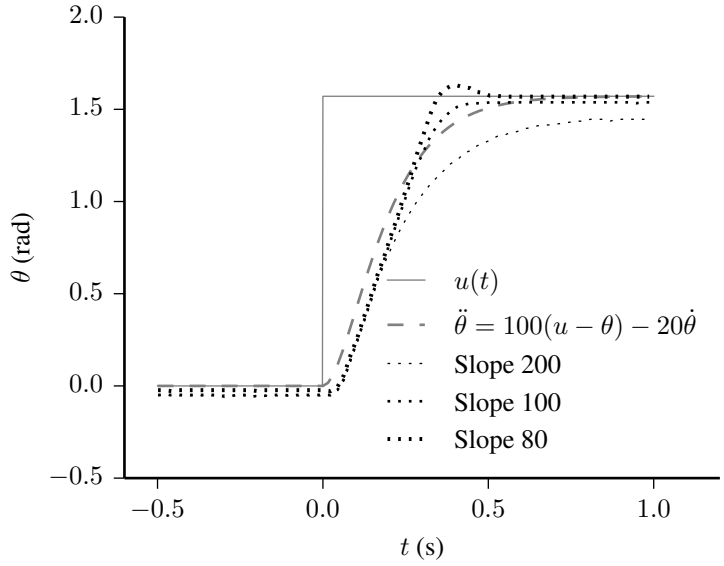


Figure 9.7: Step response of the AX-12 servo. Three different slope parameters are used (80, 150, and 200, dotted lines of increasing thickness). For comparison, the control signal $u(t)$ is also shown (solid line), along with the response of a critically damped linear system (dashed line).

parameter, we see an under-damped response which overshoots the target value. For comparison, we also show the step response of the critically damped linear differential equation

$$\ddot{\theta} = 100(u - \theta) - 20\dot{\theta} \tag{9.1}$$

which would be close in form to the expected step response if the negative feedback was exactly linear and remaining properties of the system amounted to critical linear damping (note however that the parameters of this comparison system have not been chosen specifically to model the system under study, merely to illustrate a comparable critical damping curve).

In the simulation, we use a rotational spring joint supplied by the Bullet Physics engine. This is parametrised by a damping value and a stiffness parameter to the spring. We use the control signal $u(t)$ to set the equilibrium point of the simulated spring. Setting the stiffness parameter to 100 and the damping parameter to 2.5×10^{-4} yields the step response in figure 9.8. Again this is shown compared to the idealised linear system – the simulation as might be expected is much closer to this.

This section has described the physical characteristics of the servos, and how they have been modelled in the simulation. In general the simulation has been designed to behave in a closely comparable way to the real robot, but is by no means exact. However, it is accurate enough that the gait generation algorithms described in the next section behave remarkably similarly in both simulation and the real robot. This is significant since, as we will see, the algorithm depends

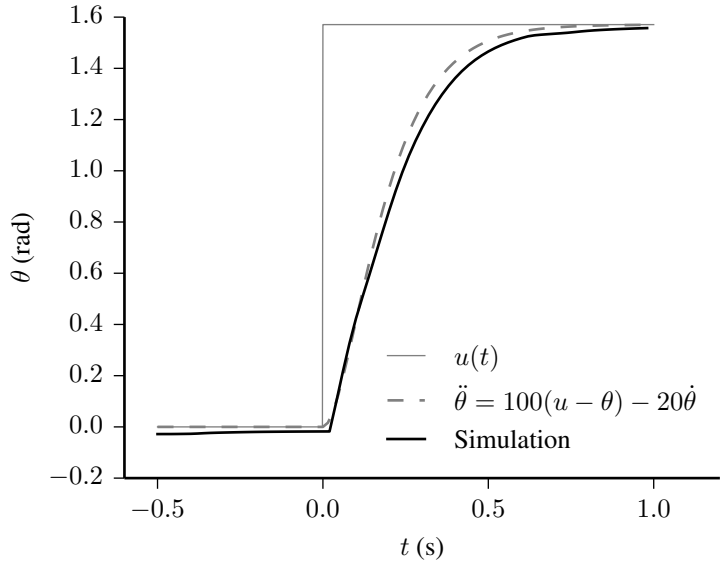


Figure 9.8: Motor step response in simulation. The set point is initialized at 0 and set to $\pi/2$ at $t = 0$, the solid black line shows the joint angle reported by the simulation over time, the solid gray line shows the step input, and the dashed grey line shows the idealised linear system.

critically on using the position sensors of the servos to obtain information about the dynamical state of the robot.

9.3 Gait algorithm

This section describes the distributed gait generation algorithm. At this stage, the simulation and real robot present identical models from the point of view of the controller, and thus what follows applies equally in both contexts. There are two primary components to the gait generation approach. First, each leg is connected to one of six identical oscillation generators – the outputs to the servos are determined by the state of each generator. Second, a set of inter-leg coupling rules introduce weak coupling between the oscillation generators and thus cause the legs to synchronise into well formed gait patterns.

Recall from figure 9.3 that each leg has three angles α, β and γ and that where necessary we distinguish the angles for each leg by subscripts α_i for leg i etc. Each angle is controlled by a single servo as discussed in section 9.2.1, the control signal (symbol u in the previous discussion) for a given joint will be referenced with a hat symbol: $\hat{\alpha}$ is the control signal for a coxa joint, for example. In general, where the angles α, β, γ appear as parts of the control algorithm described below, they refer to the angles as measured by the servo mechanisms internal sensor or reported by the simulation.

9.3.1 Oscillation generators

The motion of each leg is a series of step cycles, where each cycle is broken into two phases – the *stance* phase, where the foot is in contact with the ground, pushing the body forward, and the *swing* phase, where the foot is raised of the ground and returns to the starting position for the stance phase. The point where the stance to swing transition occurs is referred to as the posterior extreme position (PEP), and the swing to stance transition occurs at the anterior extreme position (AEP).

The the oscillation generator thus has two states, stance and swing, and switches between them at the AEP and PEP. The AEP and PEP are in turn defined by coxa angles – when α reaches $\alpha_{AEP} = 15^\circ$ in swing mode, the AEP is reached and the oscillator switches to stance mode. Similarly, define the PEP by $\alpha_{PEP} = -15^\circ$, at which point an oscillator in stance mode will switch to swing mode.

The simplest control outputs to define are the femur and tibia joints. In swing mode, the foot should be raised off the ground, and in stance mode the foot should exert pressure on the ground to support the body. For the time being, the femur and tibia joints are operated according to a simple kinematic scheme. In stance mode, we set $\hat{\beta}_{st} \leftarrow 10^\circ$ – this is supplied at the control signal u to the the femur joint. The corresponding tibia angle is $\hat{\gamma}_{st} \leftarrow 100^\circ$ – this results in the foot assembly being oriented vertically downwards. For swing mode, we need to raise the foot off the ground, achieved by increasing the femur angle. Thus we set a lift angle $\theta_l = 20^\circ$, and set the swing mode values

$$\begin{aligned} \hat{\beta}_{sw} &\leftarrow \hat{\beta}_{st} + \theta_l \\ \hat{\gamma}_{sw} &\leftarrow \hat{\gamma}_{st} + \theta_l \end{aligned}$$

Increasing $\hat{\beta}$ in this way lifts the foot off the ground. Increasing $\hat{\gamma}$ by the same amount slightly lowers the foot, but tends to keep the foot assembly oriented vertically, which improves the stability of the robot.

Though the control of the femur and tibia joints is in a sense purely kinematic, they still allow some compliance. This means that when the feet are on the ground, the actual angles of the joints will deviate slightly from their target points. For example, the coxa joint will tend to bend in a positive direction due to the weight of the robot, and the tibia joint will be somewhat compliant to lateral forces on the robot. This "passive" compliance (the deviations are not fed back to the main control program for these joints) means that these joints behave in a less "rigid" manner, but the net effect is still that the foot is lowered on stance and raised on swing.

The control of the coxa joint α is substantially more involved. First because these make greater

use of joint feedback in generating control signals, and second because this is where the inter-leg coupling will be introduced. However, for the present discussion we will describe the control system in the absence of coupling.

During swing, α generally increases to move the foot towards the AEP, and during stance, α decreases to bring the foot back towards the PEP. For each mode, we define a speed parameter ω_{st} (for stance) and ω_{sw} for swing. These will have opposite sign (with ω_{st} negative) and in general may have different magnitudes depending on the gait being generated. The $\hat{\alpha}$ set point is determined in stance mode as

$$\hat{\alpha}(t) \leftarrow \alpha(t) + k_p\omega_{st} + k_i \sum_{T=t_{AEP}}^{t} (\omega_{st} - \dot{\alpha}(T)) \tag{9.2}$$

Where k_p and k_i are two chosen constant (positive) parameters where suitable values can be determined heuristically or by trial and error. Assume that time t is measured relative to the number of control loops, i.e. in our case the main control loop runs at 50Hz – measurements are taken and set points generated at this frequency, and t can be viewed as an integral count of the number of control cycles.

The intuition behind this is as follows. Recall that the joint error $\alpha - \hat{\alpha}$ is approximately inversely proportional to the torque applied by the armature, as described in section 9.2.1. Thus with $k_p = k_i = 0$, we would generally expect no additional torque to be generated by the servo – i.e. if the output shaft was moved by external forces the controller would not counteract those forces. Thus the first term on the right hand side of equation 9.2 is a positive feedback term which incorporates the external forces applied to the joint into the future position of that joint.

The second term, $k_p\omega_{st}$ introduces a fixed force in the direction of the desired movement. Note that there is no feedback in this term (positive or negative) and thus this does not affect how the joint reacts to external perturbations. Furthermore, partly due to the lack of feedback, it is difficult to predict how this will affect the actual movement of the output shaft.

The more predictable term is the last one: $k_i \sum_{T=t_{AEP}}^{t} (\omega_{st} - \dot{\alpha}(T))$. This introduces integral negative feedback on the angular velocity of the output shaft. The measured velocity of the output shaft is estimated naively by the controller from its record of the position signal, i.e. simply as

$$\dot{\alpha(t)} \leftarrow \frac{\alpha(t) - \alpha(t-1)}{\Delta t} \tag{9.3}$$

with $\Delta t = 0.02\text{s}$ the time interval between control loops. Thus a proportion of the cumulative sum over the stance period (i.e. between t_{AEP} – the time of the most recent AEP – and the current time t) of the velocity error $\dot{\alpha}(t) - \omega_{st}$ is subtracted from the armature force. This gives a delayed response to any perturbations or errors in the velocity of the coxa joint – thus the total

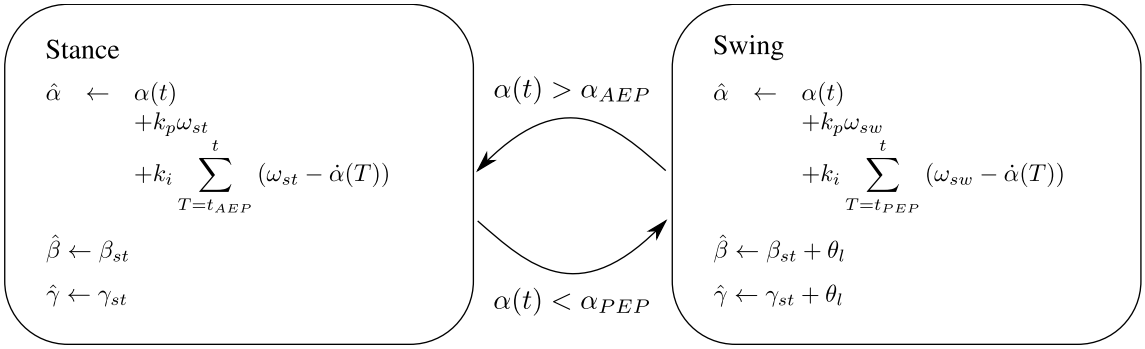


Figure 9.9: Oscillation generator. The algorithm has two states, *swing* and *stance*, transitions occur when the coxa angle reaches AEP or PEP as shown. The only input used at each time step t is the coxa angle $\alpha(t)$. The output control signals for each joint are calculated as shown and described in the text.

effect of the control system is that any external perturbation will initially be allowed to affect the joint passively, and accelerate the servo horn to match the external forcing, but after a time period (depending on the magnitude of the parameter k_i), the integral term will accumulate such that the coxa servo will tend to counteract these external perturbations. This amounts to a type of limited positive feedback comparable to the local positive velocity feedback architecture used in some versions of the Walknet controller (Cruse et al., 2002, 2007) – the first term introduces positive feedback, and the third term introduces time delayed negative feedback. This allows the legs to transfer information mechanically by exerting forces on each other. This is particularly relevant when multiple feet are in contact with the ground (i.e. in stance mode) – those feet which are in contact with the ground at the same time will exert forces on each other whenever they exert forces against the closed kinematic chain formed by themselves, the robot's body and the ground.

It is critical for this that the negative feedback is based on velocity and does not correct the absolute position of the joint – that is, if the joint is for example, "slowed down" from it's normal course, it will eventually return to its target speed, but it will not attempt the counteract the phase lag that was introduced by the temporary external force.

This means that even without the explicit coupling rules which we come to in the next section, the embodiment of the robot introduces mechanical coupling between the otherwise separate oscillation generators. This can act to change the phase differences between legs and introduce synchronisation, as will be demonstrated in section 9.5.

For clarity, an overall description of the oscillation generator is shown in figure 9.9. The key elements are the two states (stance and swing), the transition rules between those states (AEP and PEP), and the algorithms for calculating the joint set points.

9.3.2 Coupling rules

A copy of the oscillation generator described above is set up for each leg. In order to generate a useful gait, we need the oscillators to synchronise in such a way that the legs move in a globally coordinated manner. We achieve this by implementing a set of coupling rules described in this section. This is inspired by the coupling rules used in Walknet (Cruse et al., 2007), of which there are six. However, the rules used by Walknet are rather more specific than ours, which are as follows:

1. Swinging legs suppress the stance-swing transition in the posterior leg. When a leg reaches its PEP, it will always wait until the leg in front (i.e. the leg immediately anterior on the same side) is in stance mode before transitioning.

2. When a leg reaches AEP, the immediately posterior ipsilateral leg should move towards its PEP.

3. When a leg reaches AEP, the opposite contralateral leg should move towards its PEP.

Rule 1 is comparable to rule 1 in Walknet (Cruse et al., 2007), though it operates in the opposite direction (anterior legs suppress posterior in our model, in Walknet, posterior legs suppress anterior). This rule helps to ensure static stability since it reduces the change of two neighbouring ipsilateral legs swinging at the same time – if they do, then the robot is likely to "tip" towards the pair of swinging legs, resulting in them contacting the ground when they are in swing mode.

Our rules 2 and 3 compare to rule 2 of Walknet, but although they operate in the same manner, we distinguish them here since in some experiments we will not always use rule 3. These rules tend to to induce anti-phase synchronisation between neighbouring legs. Walknet has four further rules which we do not adopt analogues of. Thus our model is substantially simpler than Walknet, but as we will see is sufficient for successful gait generation.

The three rules all act "locally" in that they only affect one of their immediate spatial neighbours, either ipsilaterally or contralaterally – the combination of all the rules (see figure 9.10) ensures that there is potential for global synchronisation of all of the legs in order to introduce a gait pattern.

The implementation of rule 1 is straightforward and rigid compared to the approach used for rules 2 and 3 which we discuss below. For rule 1, when a leg reaches its PEP in stance mode, the controller checks whether the anterior leg is swinging, and if so, the stance-swing transition in the posterior leg does not occur. Additionally, the coxa target position in the posterior leg is set to $\hat{\alpha} \leftarrow \alpha_{PEP}$ – this replaces the ordinary calculation for the coxa joint and so the coxa will tend to

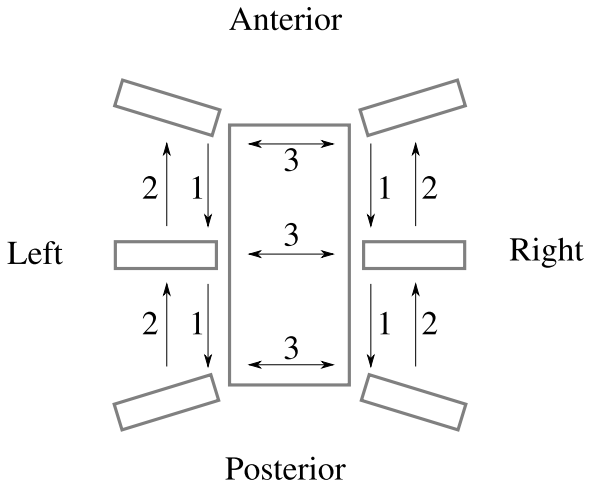


Figure 9.10: The structure of the three coupling rules. The directions in which legs interact with each other are shown by the arrows labelled by the rule they correspond to: 1 – suppression of swinging in posterior leg when anterior leg is swinging; 2 – matching of posterior PEP to anterior AEP; 3 – matching of contralateral PEP to anterior AEP.

remain at the PEP angle until the anterior leg has finished its swing. Once the anterior leg finishes its swing, the posterior leg is then able to transition to swing mode itself.

Rules 2 and 3 are implemented using a form of pulse-coupling through a set of "input currents" for each leg, $I_i(t), i \in \{0, 1, \ldots, 5\}$ which we now describe. The currents will be set according to phase differences between legs when any leg reaches its AEP, the currents modulate internal excitation variables associated with each leg, and the leg target speed is updated according to the value of the excitation variable.

Define, according to the leg layout in figure 9.2, $OPP(i)$ as the leg opposite leg i (e.g. $OPP(0) = 1$, $OPP(1) = 0$) and $POST(i)$ as the leg immediately posterior to leg i (e.g. $POST(2) = 0$, note that $POST(0)$ and $POST(1)$ are undefined since there are no legs posterior to the two back legs). Define the normalised distance $D(i)$ of a leg i from its PEP as

$$D_i(t) = \frac{\alpha_i(t) - \alpha_{PEP}}{\alpha_{AEP} - \alpha_{PEP}}$$

Note that due to external forces or over activation is it possible to have $D < 0$ or $D > 1$ since the coxa joint will generally slightly "overshoot" the PEP and AEP angles before switching direction to a small degree. However, the above formula approximately gives a measure of the leg's distance from its PEP from 0 (leg is at PEP) to 1 (leg is at AEP).

Now whenever a given leg i reaches its AEP, we can define a relative phase advancement variable for any other leg $j \neq i$:

For any leg i which has reached its AEP (i.e. where $\alpha_i(t) > \alpha_{AEP}$ and we are about to undergo the swing-stance transition):

1. Apply rule 2 if $i \geq 2$ (i.e. $POST(i)$ is defined):
   (a) $I_i(t) \leftarrow I_i(t) + \phi_{POST(i)}$
   (b) $I_{POST(i)}(t) \leftarrow I_{POST(i)}(t) - \phi_{POST(i)}$
2. Apply rule 3:
   (a) $I_i(t) \leftarrow I_i(t) + \phi_{OPP(i)}$
   (b) $I_{OPP(i)}(t) \leftarrow I_{OPP(i)}(t) - \phi_{OPP(i)}$

Table 9.1: Input current calculations.

$$\phi_j(t) = \begin{cases} \max(0, D_j(t)) & \text{leg } j \text{ is in swing mode} \\ -\max(0, D_j(t)) & \text{leg } j \text{ is in stance mode} \end{cases}$$

Thus ϕ_j will be at -1 when leg j starts its stance phase at its AEP, progresses to 0 as it reaches its PEP and continues increasing to 1 as it approaches its AEP in swing mode.

At each time step t, the input currents $I_i(t)$ are set to zero and then updated by the algorithm shown in table 9.1.

The input currents are in turn used to modify an internal "excitation" variable variable $\zeta_i, i \in \{0, 1, \ldots, 5\}$, again introduced for each leg. This variable is updated on each control loop according to

$$\zeta_i(t) \leftarrow \tau_i(t)\zeta_i(t-1) + I_i(t)$$

where τ_i is a decay constant < 1 and $I_i(t)$ is the input current which introduces the coupling signals from the other legs in accordance with rules 2 and 3. The decay constant ensures that ζ_i tends to zero over time in the absence of inputs from the other legs. Since τ_i effectively sets a time period over which the excitation signal decays, it is adjusted relative to the overall expected time period of the current phase of the associated leg. That is, we choose a constant parameter to the algorithm, k_τ and set

$$\tau_i(t) \leftarrow \begin{cases} {k_\tau}^{|\omega_{st}|} & \text{leg } i \text{ in stance} \\ {k_\tau}^{|\omega_{sw}|} & \text{leg } i \text{ in swing} \end{cases}$$

Choosing $0 < k_\tau < 1$, this results in a faster decay of ζ_i the higher the target speed of the current leg. We additionally ensure that the coupling influence cannot persist for a large proportion of the step cycle by setting $\zeta_i = 0$ whenever leg i transitions from stance to swing or swing to stance mode.

The value of ζ_i then influences the behaviour of the coxa joint for leg i by modifying the actual target speed used by that leg. Thus define a new target speed variable which is initially the standard stance or swing target speed:

$$\omega_i(t) \leftarrow \begin{cases} \omega_{st} & \text{leg } i \text{ in stance} \\ \omega_{sw} & \text{leg } i \text{ in swing} \end{cases}$$

This is then updated by multiplicatively by the excitation variable:

$$\omega_i(t) \leftarrow \max(0, 1 + k_c\zeta_i(t))\omega_i(t)$$

Another constant k_c affects the overall strength of the coupling for rules 2 and 3. This new ω_i is used in place of ω_{st} or ω_{sw} wherever it appears in figure 9.9 in the oscillation generator. That is, positive values of ζ_i increase the magnitude of the target speed used to drive the coxa joint set points.

The intuition behind this can be seen as follows, using rule 2 as an example. The rule encourages a pair of legs i and $POST(i)$ to synchronise in anti-phase such that when i reaches its AEP, $POST(i)$ will reach its PEP. Thus, when leg i reaches AEP, if $POST(i)$ is currently behind its PEP (i.e. in stance), it will speed up (because positive current will be added to $I_{POST(i)}$, subsequently increasing $\zeta_{POST(i)}$ which in turn amplifies the magnitude of the target speed of $POST(i)$). The the leg $POST(i)$ will reach its PEP sooner as a result of this coupling. Simultaneously, leg i will slow down so that it reaches its next AEP later (negative current is added to I_i). Conversely, if $POST(i)$ it advanced beyond its PEP (i.e. in swing mode), it will slow down while leg i speeds up. The amount of slow down or speed up in each case is proportional to how far delayed or advanced $POST(i)$ currently is, so when the legs are already synchronised in anti-phase, the effect will be negligible.

Similar logic for rule 3 applies in respect of the leg opposite. The internal excitation variables ζ_i are needed to "smooth out" the pulse-coupled influences, because each leg has a delayed response to changes in its target speed, partly due to the inertia of the leg and partly due to the integral component of the oscillation generator. Without this internal smoothing, the coupling pulses are either too small to have any effect, or (if the coupling constant is set very high) simply induce very "jerky" movement in the legs.

Parameter	Value
k_p	0.04
k_i	0.0025
ω_{sw}	2
ω_{st}	-2
k_c	2
k_τ	0.9

Table 9.2: Parameter values for coupling demonstration

The synchronisation process is illustrated in figure 9.11, which shows traces of the coxa angle and excitation variable at the start of a brief experiment in simulation. In this experiment, the robot is placed on top of a cuboid such that its legs are free to move without touching the ground (such that coupling in this instance occurs only via the internal coupling rules, and not through mechanical contact with the ground). The parameters are set to the values shown in table 9.2. The plot shows traces taken from the three right hand legs, in order to illustrate the effect of rule 2. For further clarity, rule 1 (stance-swing suppression) and rule 3 (contralateral coupling) are disabled in this experiment – thus only rule 2 is shown.

At the beginning of the traces (in the first second), the coxa joints move to randomly chosen starting positions – this is before any of the gait generation algorithm is started. After one second the algorithm begins, with all legs starting in stance mode. Thus due to the choice of initial position, the initial phase of each leg differs. In this case, we have leg 1 slightly delayed relative to leg 3, which in turn is also slightly delayed from leg 5. Coupling takes effect when each leg reaches AEP. When leg 5 reaches its first AEP at $t = 1.94$, leg 3 is in swing mode, therefore a negative input pulse is added to leg 3 (such that its swing cycle will be delayed) and a positive input pule is added to leg 5 (which advances its swing cycle). When leg 3 reaches its first AEP at $t = 2.44$, a pulse is generated affecting leg 1 – since leg 1 is in its swing phase at this point, a negative pulse is added to ζ_1 (thus slowing down the leg 1 coxa joint α_1) and a positive pulse is added to ζ_3. As a result, over time, leg 5 receives positive pulses which tend to advance its cycle, and leg 1 receives negative pulses and its phase is delayed. Since leg 3 receives pulses in both directions, its phase is not adjusted as much.

Over time, this leads the three legs plotted in figure 9.11 to synchronise in anti-phase with the adjacent leg – once they have done so, each leg reaches AEP when the leg behind is close to PEP, and so the strength of the pulses (which is dictated by the distance from PEP) decays. Thus after

reaching synchrony, the coxa joint speeds are not affected by coupling as strongly.

9.4 Generating gait patterns

Thus above discussion concludes the description of the control algorithm. In this section, we present a number of simulation and real robot experiments in which we demonstrate the ability of this algorithm to generate multiple gaits. We can also demonstrate conclusively that the mechanical coupling between legs, as a result of their simultaneous contact with the ground, has a significant impact on the emergence of synchrony between the legs.

To begin with, we show how the gait algorithm can be used to generate multiple gaits. Conceptually, the approach is very similar to Walknet as well as CPG approaches. There are two principle factors that vary between gaits:

1. The stance/swing ratio $r = T_{st}/T_{sw}$ (the ratio of time spent in stance mode T_{st} to time spent in swing mode T_{sw}). This can be thought of as the overall speed of the gait – keeping T_{sw} fixed, a higher ratio will lead to longer stance times and hence fewer complete step cycles in a given period.

2. The phase pattern of the legs under synchrony. Different gaits admit different symmetry patterns between the legs.

Our algorithm, like many CPG approaches, allows one to vary the gait by modifying only the stance/swing ratio or speed parameter r. The correct phase pattern emerges as a result of the coordination between the legs created by the internal coupling rules and the embodied mechanical interactions in the robot.

9.4.1 Tripod and metachronal gaits

We begin by investigating two common hexapod gaits. The first gait is the tripod gait, which is typically the fastest gait, and corresponds to a ratio $r = 1$ – that is, the swing and stance phases have the same duration. There are two statically stable tripod configurations where three of the robots legs are on the ground – the middle leg on one side, and the front and back legs on the other. The tripod gait simply switches between these two stable configurations, with each having equal duration. The phase pattern can be written as a vector

$$\mathbf{\Phi} = (0, \pi, \pi, 0, 0, \pi)$$

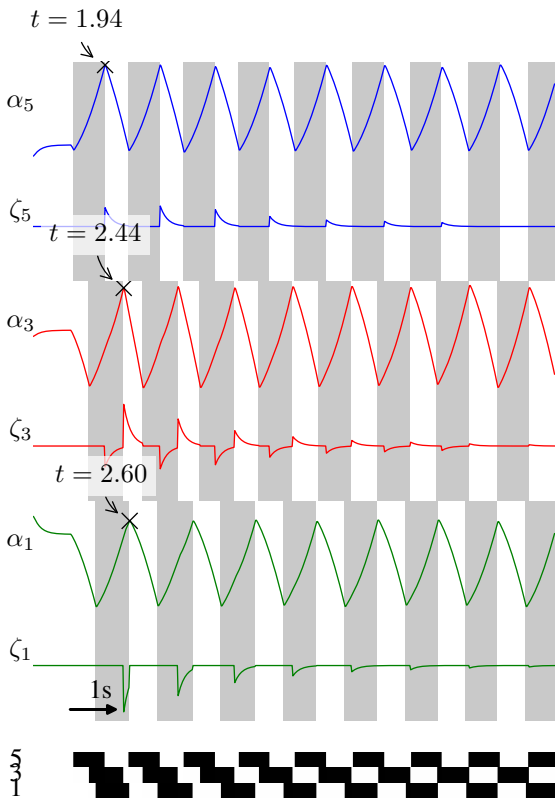


Figure 9.11: Operation of rule 2 coupling in simulation. Traces are shown for each of the three right hand side legs (legs 1, 3, and 5, see numbering in figure 9.2). For each leg, the coxa joint angle α and internal excitation variable ζ are plotted. Additionally, the traces for each leg have a grey background where the oscillator is in swing mode. The bottom plot shows a raster plot of the three legs, with black indicating swing mode and white indicating stance mode. The first AEP of each leg is indicated with a cross.

where the element Φ_i is the relative phase of the ith leg from the 0th leg (according to the numbering in figure 9.2).

The second gait we consider is the metachronal gait. This equates to slow walking, with the ratio $r = 3$ – thus each leg spends only one quarter of the time in swing mode. Here there are four statically stable leg configurations employed – two have just one of the middle legs in swing mode, and the rest of the legs in stance, and the other two have diagonally opposite front and back legs in swing mode. The pattern vector is

$$\mathbf{\Phi} = (0, \pi, \frac{3\pi}{2}, \frac{\pi}{2}, \pi, 0)$$

We adjust the gait by altering the ratio $\rho = |\omega_{sw}|/|\omega_{st}|$ – specifically, we keep ω_{sw} fixed and change ω_{st} to give the desired ratio. In principle, since ρ is the ratio of angular speeds in swing and stance modes, it should closely correspond to the ratio of time spent in each mode r. In practice, the actual speed at which the coxa joints move is affected by the mechanical properties of the robot and the control system we use. Thus to achieve different gaits, we simply find by trial and error values of ρ which give the desired r – for the tripod gait, $\rho = 1$ corresponds to $r = 1$, for the metachronal gait, we used $\rho = 4.0$ to get $r = 3$.

In order to assess the performance of the algorithm, we construct a heuristic measure of the closeness of a given time series from the robot to a desired gait (specified by its pattern vector $\mathbf{\Phi}$) as follows.

First, synchrony in any case requires that all legs oscillate as the same fundamental frequency, since our gaits return globally to the same point after any given leg completes a step cycle. Thus the first component of our measure is based on the variance of the fundamental frequencies of the leg oscillations. We find this by taking the discrete Fourier transform of the coxa joint position measurement α_i from each leg and finding the frequency of the first peak in amplitude, giving a fundamental frequency f_i for each leg i.

However, we also require, importantly, that the legs achieve the desired phase offset pattern $\mathbf{\Phi}$. The phase of a given frequency component can also be taken from the existing Fourier transform – we use the phase at the fundamental frequency just found and label it ϕ_i. We find the phase difference from the desired phase according to the gait specification:

$$\Delta\phi_i = \phi_i - \Phi_i$$

If the robot is close to the correct gait pattern, the $\Delta\phi_i$ should all represent similar angles. That is, when the robot achieves the correct gait, all the $\Delta\phi_i$ should be close to the angular mean of the group:

Parameter	Value
k_p	0.04
k_i	0.005
ω_{sw}	2
ω_{st}	-2/ρ
k_c	3
k_τ	0.9

Table 9.3: Parameters for gait example experiments. These parameters were used in simulation to generate the data for figures 9.12 and 9.13. The parameter ρ is 1.0 for the tripod gait and 4.0 for the metachronal gait.

$$\bar{\phi} = \angle\left[\langle \exp(j\Delta\phi_i)\rangle\right]$$

where $j = \sqrt{-1}$, angle brackets $\langle\cdot\rangle$ represent the arithmetic mean across the leg index i, and $\angle[a + jb] = \arctan(b/a)$.

We now construct a measure designed to approach unity when the robot is close to the desired gait pattern and decrease whenever it is not. We do this by taking an inverse square exponential of the sample variance of the frequencies and multiplying by the average cosine of the difference between the phases and their average:

$$\sigma = \exp(-\mathrm{var}(f_i)^2)\left\langle \cos(\Delta\phi_i - \bar{\phi})\right\rangle \tag{9.4}$$

Both terms approach unity under the "ideal" synchronisation conditions of zero variance in f_i and all the $\Delta\phi_i$ take similar values.

This measure can be used to evaluate the performance of the algorithm in generating different gaits. Using the parameters in table 9.3, the simulation was run for 200s under two conditions: $\rho = 1$ (tripod gait, figure 9.12) and $\rho = 4.0$ (metachronal gait, figure 9.13). These results show how the algorithm can produce either gait on modification of only a single parameter ρ. The starting positions of the coxa joints are initialised at random, and so each time the dynamical coupling has to bring their motion into the desired pattern. This can be seen in the raster plots showing the process of synchronisation from different starting points in the bottom left of figures 9.12 and 9.13.

The tripod gait is realised very quickly, appearing after only a few seconds as indicated by the plot σ at the top of figure 9.12. The raster plots on the bottom left of figure 9.12 also show

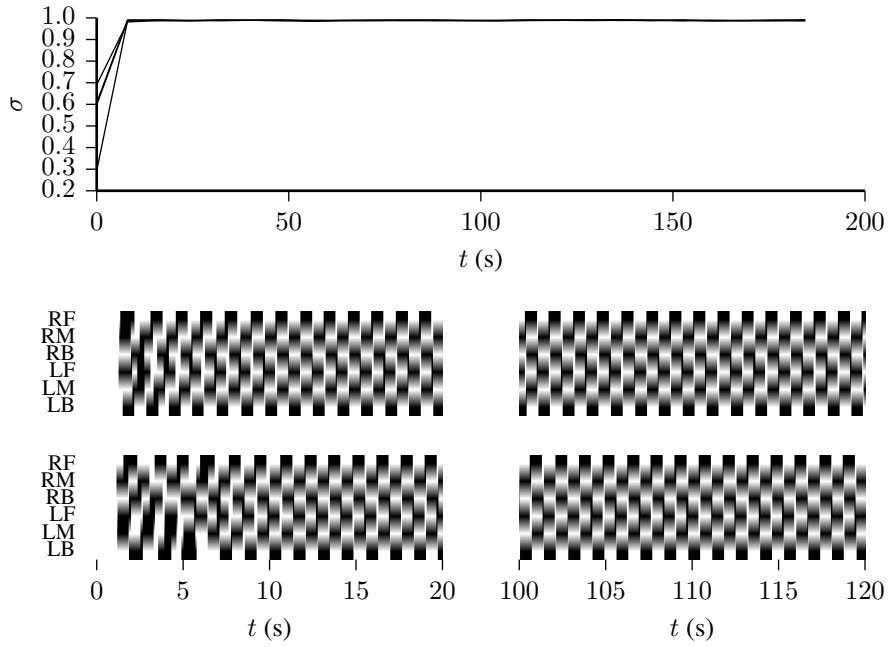


Figure 9.12: Simulation of tripod gait. The top graph shows the synchrony measure σ applied to 8 second windows of the coxa joint angle data for four separate runs of the system. The raster plots are examples from two of the four runs, the left-hand two showing when each leg is in swing (black) or stance (white) in the first 20s, the right-hand pair show the time period between 100s and 120s. In the raster plots the legs are labelled LB = left back (posterior), LM = left middle, LF = left front (anterior), RB = right back, RM = right middle, RF = right front. These correspond respectively to indices 0, 2, 4, 1, 3, 5 in figure 9.2.

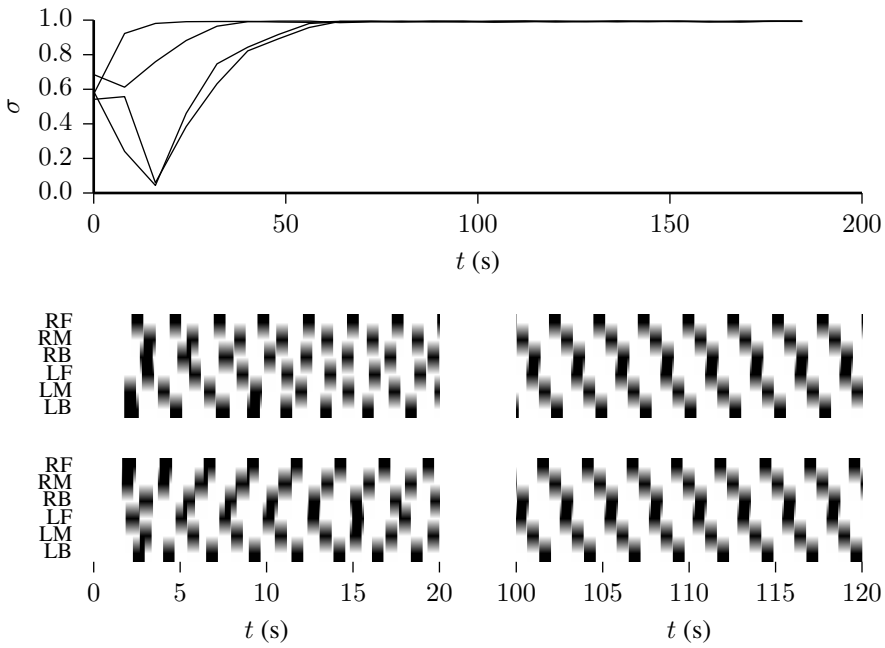
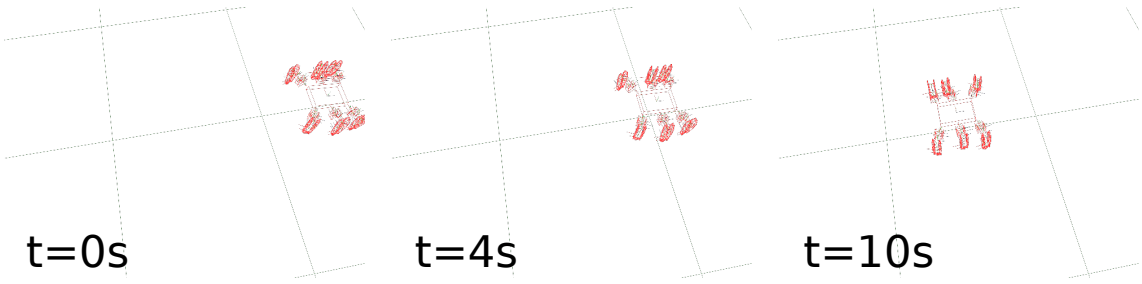


Figure 9.13: Simulation of metachronal gait. The plots show the same analysis as in figure 9.12 but for the swing-stance ratio $r = 4$ (metachronal gait), achieved by setting the parameter ρ to 4.0. This time, it takes substantially longer for the more complex gait to emerge.

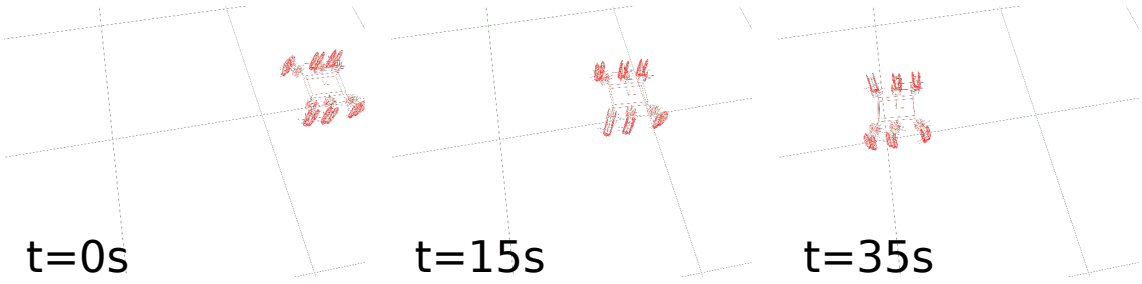
the rapid emergence of the desired gait pattern. In figure 9.13 we see that the metachronal gait emerges much more slowly – the raster plots do not show a stable metachronal gait after the first 20s, but we see that much later (at 100s) the gait has stabilised. The slow stabilisation is also indicated by the relatively slow convergence of σ to unity in the metachronal experiment. One possible reason for the slower convergence is that the overall step cycle is twice as long for the metachronal gait, meaning that synchronisation pulses occur less frequently. This alone does not account for the large discrepancy, which may also be due to less quantifiable factors arising from the relative complexity of the metachronal gait.

9.5 Replacing internal coupling with mechanical coupling

We have seen the algorithm can generate gaits successfully, however we have no clear evidence of the role of the *mechanical* coupling in gait generation. Here we will show the first such evidence that arises when we disable rule 3 – the contralateral coupling rule. Recall that rule 3 encourages legs on opposite sides of the robot to synchronise in anti-phase. Note also from figure 9.10 that without rule three, the two sets of legs on either side form distinct, uncoupled, groups. Thus the behaviour of the robot after removing rule 3, particularly with respect to the synchronisation between the two (internally uncoupled) groups of legs will exemplify the effect of the mechanical



(a) Tripod (no rule 3)



(b) Metachronal gait

Figure 9.14: Video frames of simulated robot walking. Screen shots taken from the simulator physics drawing. The three frames in (a) are from a run using the tripod gait with no contralateral coupling (rule 3 disabled), the frames in (b) show the metachronal gait.

coupling. Here we will see that even without rule 3, a stable tripod gait and be generated, demonstrating that mechanical coupling can significantly alter the behaviour of the system. In the next chapter, we will continue this analysis by examining the effect of eliminating rule 3 on transfer entropy between legs.

To illustrate the behaviour of the robot in simulation, screenshots of the simulation output are shown in figure 9.14 – we can see here that even with rule 3 disabled, the robot successfully walks across the visual field. This figure also shows that the tripod gait is substantially faster in terms of the overall speed that the robot moves along the ground than the metachronal gait.

Maintaining the parameters as they are in table 9.3 (with $\rho = 1$), even without rule 3 a successful tripod gait is generated as shown in figure 9.15, though the synchrony measure σ varies somewhat compared to figure 9.12 – i.e. compared to the same system with rule 3 enabled.

To demonstrate that the synchrony achieved here is genuinely a consequence of the mechanical coupling through the ground, we plot in 9.16 the results from the simulation where a cuboid block is used to keep the robot's feet from touching the ground. Without this mechanical contact, there is no feedback to synchronise the legs on the left and right sides of the robot.

This confirms that mechanical coupling has a significant impact on the behaviour of the robot, due to the slight compliance that was factored into the leg control scheme. It also confirms that this mechanical coupling in general acts to synchronise those legs which are in contact with the

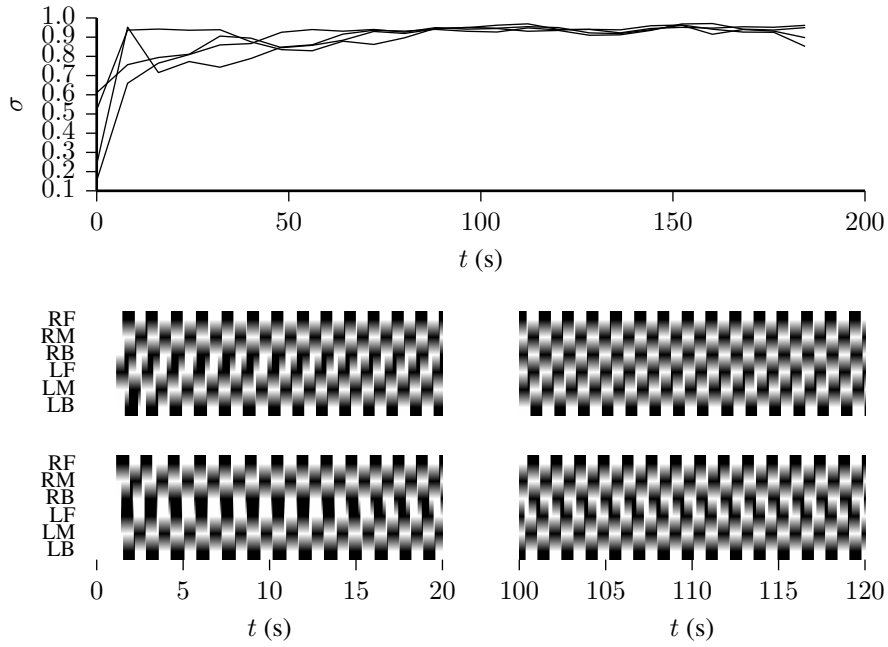


Figure 9.15: Simulation of tripod gait without rule 3. Here the swing/stance ratio parameter is $\rho = 1$ as in the tripod gait, however, rule 3 is not applied. In spite of this, successful tripod gait synchrony is achieved.

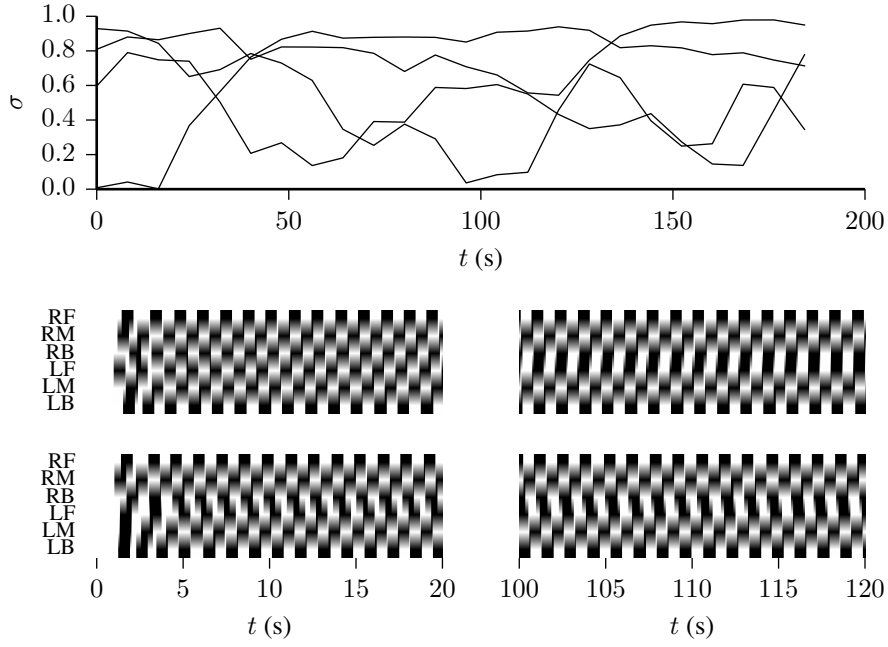


Figure 9.16: Simulation of tripod gait without rule 3, robot raised off ground. The robot is unchanged from the one used in figure 9.15, except that it is placed on a cuboid object that keeps its feet from touching the ground. Without the required mechanical coupling, gait synchrony is not reliably achieved. Note however that the synchrony between the left and right groups of three legs is maintained due to the internal coupling (via rules 1 and 2).

ground together in phase.

9.5.1 Gaits on the real robot

The gaits can also be generated by running the same algorithm on the real robot. However, due to various discrepancies between the simulation and the real system, slight adjustments must be made to the algorithm parameters to work on the real robot. Starting with the parameters used in simulation (table 9.3), the following changes were made to achieve relatively stable gaits on the real system:

- k_i – The integral control parameter is changed from 0.005 in simulation to 0.01 in the real robot. This should lead the controller to counteract errors produced by the servo more quickly – since the real servos are likely to have more complex and non-linear torque and friction characteristics than the idealised simulation, increasing this parameter slightly enabled the controller to more quickly bring the movement of the servo to its "desired" position.

- k_τ – The decay constant for the excitation variables is increased from 0.9 in simulation to 0.95 in the real robot, meaning the excitation variables decay more slowly.

- k_c – The coupling parameter was set to 3 in simulation. In the real robot we use the lower value of $k_c = 1.5$. Note that the overall "strength" of the coupling is effectively determined by both k_c and k_τ – with a lower k_c, the coupling pulses sent between the legs in the real robot are smaller, however with a higher k_τ, they last for longer, and so have longer to affect the behaviour of alternative legs. It was found in the real robot that high k_c caused "jerky" behaviour due to strong pulses, but simply reducing the k_c on its own would not lead to a stable gait. Increasing k_τ at the same time as reducing k_c effectively "smooths out" the influences between legs, while keeping their long-term effect similar.

Representative results are shown in figure 9.17. For practical reasons, the robot is only run for 60s in each scenario, and we find that although the standard tripod gait remains reasonably reliable, for the real robot is is harder to generate fully stable metachronal gaits or to achieve a perfect tripod gait without contralateral coupling. However, in all scenarios the robot can reasonably be said to walk – it successfully moves across the ground with reasonable coordination between the legs. The successful locomotion can be seen qualitatively in the sample video frames shown in figure 9.18.

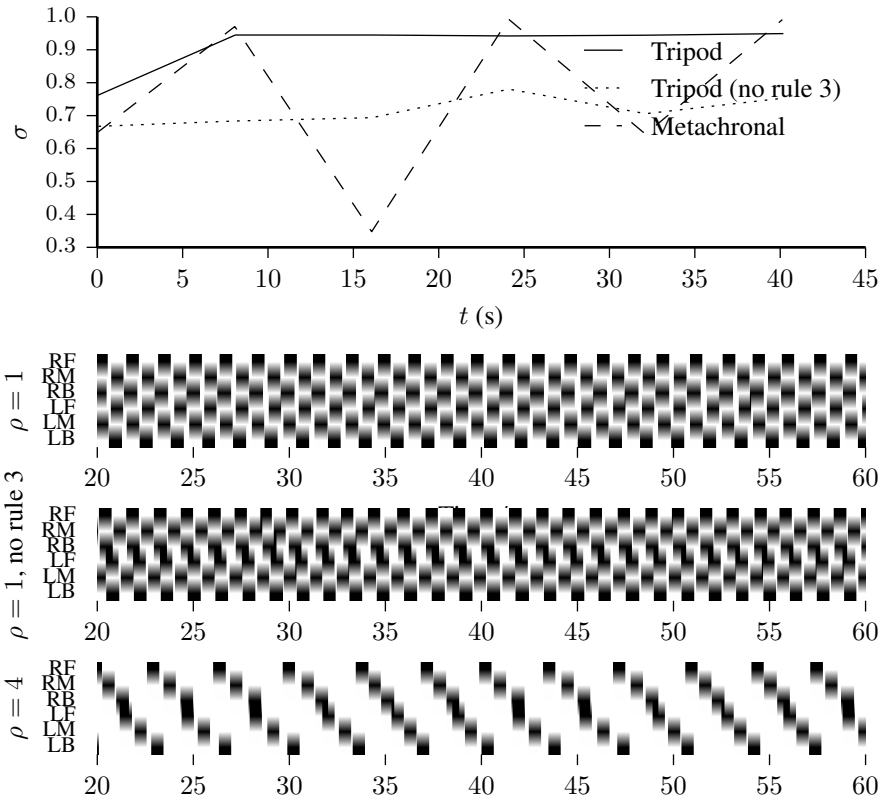
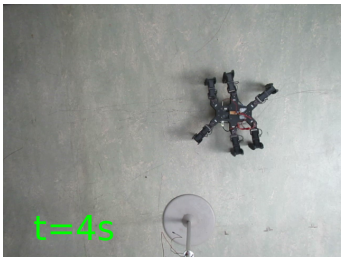
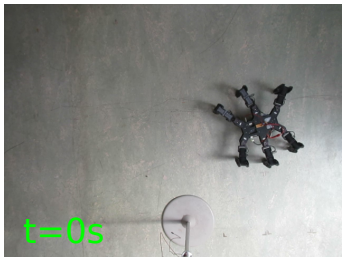
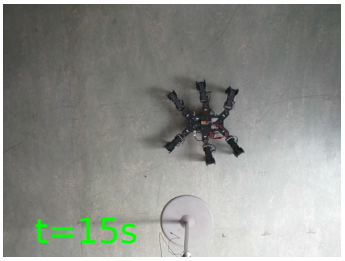


Figure 9.17: Gaits generated on the real robot. This plot shows together the results from a single run of three scenarios on the real robot: first the tripod gait ($\rho = 1.0$) with standard coupling rules, second the tripod gait where rule 3 (contralateral coupling) is disabled, and finally the metachronal gait generated with $\rho = 4.0$.



(a) Tripod (no rule 3)





(b) Metachronal gait

Figure 9.18: Video frames of real robot walking. A video camera was fixed facing downwards approximately 2m from the floor and the robot allowed to walk underneath. The three frames in (a) are from a run using the tripod gait with no contralateral coupling (rule 3 disabled), the frames in (b) show the metachronal gait. Note that both gaits successfully move across the field from right to left, but the metachronal gait is substantially slower.

9.6 Stability and robustness to noise

The examples above showed successful gait generation. The parameters were chosen to give a known good configuration. However, since the global gait pattern emerges as the result of local coupling interactions, we can expect that the overall performance of the algorithm will vary according to different parametrisations. Furthermore, the simulation should, ideally, be closely representative of the real system. In this section, we investigate the overall performance of the gait generation algorithm in simulation and introduce noise to the simulation to mimic the noisy inputs the real robot receives.

The first type of noise to consider is the quantization noise introduced by the limited resolution of the digital communication system used to send data from the servos back to the main controller. The servos measure position using a potentiometer, but the voltage from this is converted into 10-bit data packets, giving a resolution of only 1024 distinct positions that the robot can recognise within a specific range of angles. This is straightforward to model, by simply taking the angular value reported by the physics engine, and "rounding" it to the nearest quantization step that the AX-12 servo would use in its place.

To measure the effect of this change we will analyse the temporal pattern of the global syn-

chrony measure σ. Recall that values of σ close to 1 represent a good fit to the desired gait. We run the system for 300 seconds, then define the result as representing a stable gait if:

1. At some point, the value of σ is greater than 0.9; **and**

2. The median value of σ after the first time it passes 0.9 is also greater than 0.9

If so, the system has not only reached the desired gait, but has maintained it for the majority of the time after it synchronises. This allows brief periods of de-synchronisation which may occur randomly, but noes not class a gait as stable if synchrony is only reached intermittently.

We perform this analysis on 40 separate runs of the system, each time with the coxa joint angles initialised uniformly at random in the range $[-15^\circ, 15^\circ)$. The parameters used for the simulation are as shown in table 9.3 as in the gait examples above, however the coupling parameter k_c is allowed to vary, and we collect data for values of k_c between 0 and 10. Figure 9.19 shows the effect of introducing quantization noise on the stability of the gait as measured in this manner, for the tripod gait system with $\rho = 1.0$. The results show that a coupling level of approximately $k_c = 1$ is required to achieve a stable gait. However, quantization noise appears to have a relatively minor impact on the overall stability of the gait.

In addition, aside from the coxa angle measurements being at a low resolution, there is likely to be some error introduced by electrical noise affecting the potentiometer readings themselves. This is harder to model, but a simple assumption is that Gaussian noise can be added to the coxa joint angle as reported by the physics engine. This is done just before the quantization step – we expect that after the signal is converted to a digital form by the servo controller, there will be negligible errors in digital transmission.

The noise level added to the model can be varied by setting the standard deviation of the Gaussian noise, η (measured in degrees). Figure 9.20 illustrates the effect of adding Gaussian noise to the simulated coxa angle sensor with a standard deviation of η°. The values tested show a limited but clearly present effect of noise. The effect of adding noise to the system in this way is shown in figure 9.21, where again we see a modest effect on the overall stability of the gait generation algorithm, although for the highest noise level tested ($\eta = 0.4$) there is a notable range of coupling values between $k_c = 3.0$ and $k_c = 4.0$ where we see a substantial increase in the time taken to reach a stable gait.

We take the intermediate noise level, $\eta = 0.2$, as our standard for all remaining experiments in this chapter and chapter 10. Using this value, we investigate how robust the different gaits that we have previously introduce are under changes to the coupling constant k_c. Using the same experimental protocol the stability of the algorithm for the standard tripod gait, as well as the

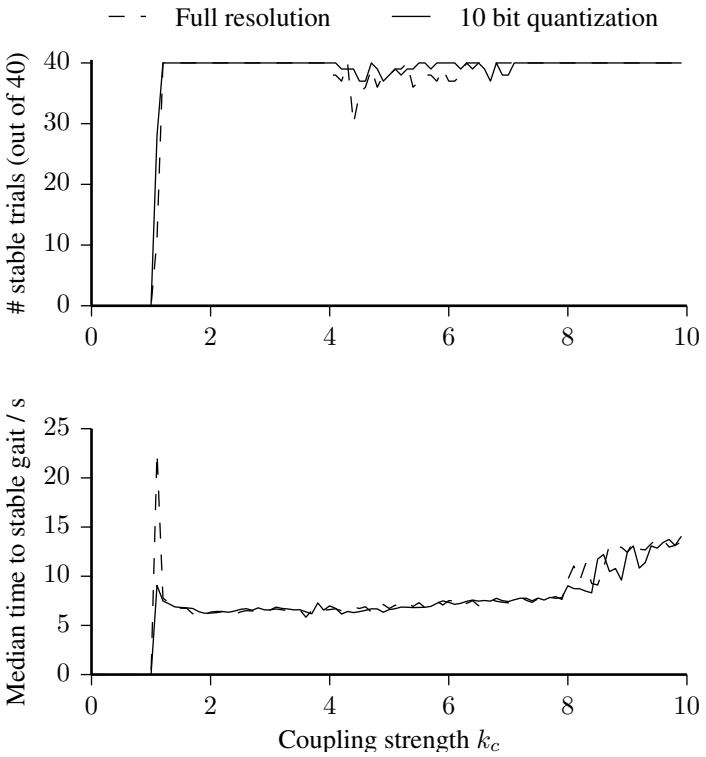


Figure 9.19: Robustness of gait generation to quantization noise. The upper plot shows the number (out of 40 trials) of times that a stable gait is achieved according to the definition in the text. The lower plot shows, for gaits which achieved stability, the median value of the first time at which the σ value passed 0.9.

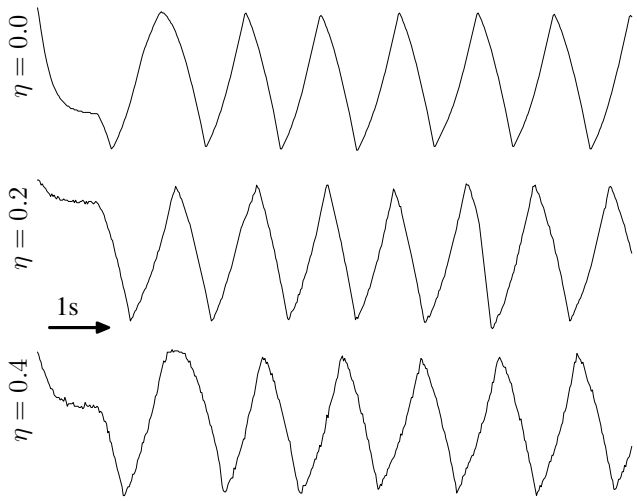


Figure 9.20: Time series traces of coxa joint angles with varying noise. The first ten seconds of data from the coxa joint angle α_0 are shown after noise and quantization has been applied. This shows the modest amount of noise that is introduced into the simulated system.

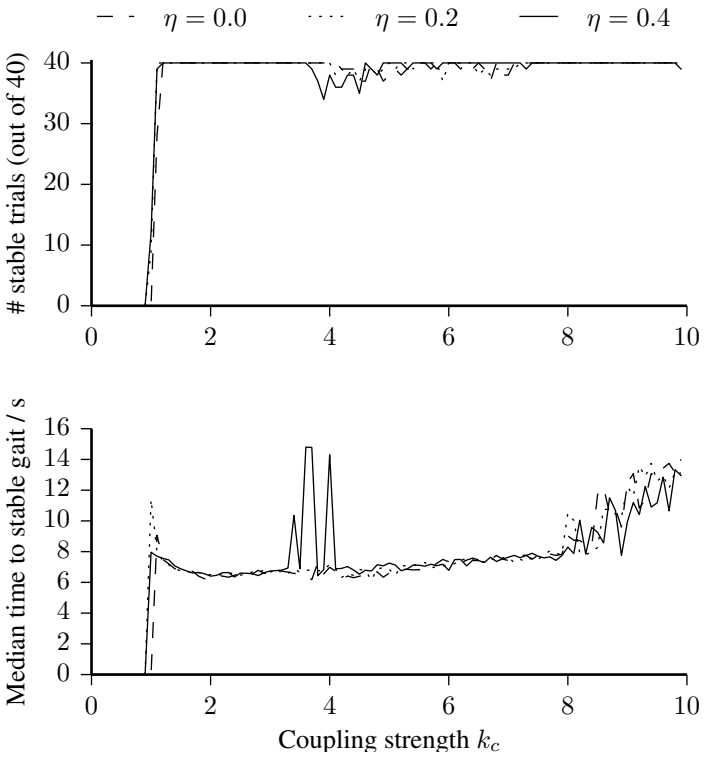


Figure 9.21: Robustness of gait generation to Gaussian sensor noise. Here Gaussian noise with the given standard deviation η is added to all joint sensor inputs before being passed to the control algorithm. Again the effects are limited, except for a noticeable reduction in the number of stable trials, and lengthened median time to reach stability, around a coupling coefficient $k_c = 4.0$.

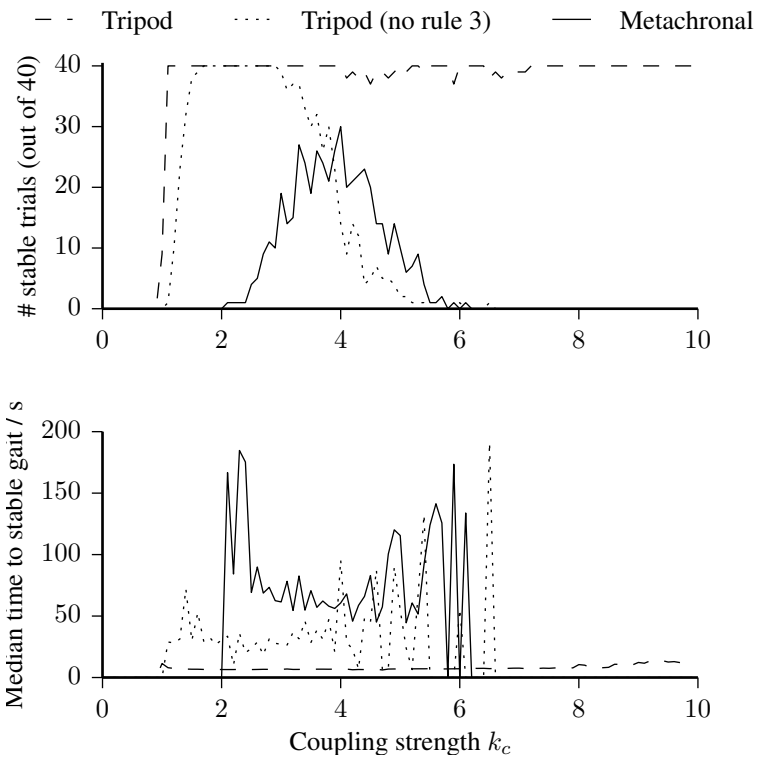


Figure 9.22: Robustness of different gait generation settings. The algorithm is set to produce a tripod gait ($\rho = 1$) as before, as well as a tripod gait without using rule 3 (i.e. no contralateral coupling), and a metachronal gait ($\rho = 4.0$).

tripod gait generated without rule 3 (i.e. without contralateral coupling) and the metachronal gait (with the stance/swing ratio parameter $\rho = 4.0$) is evaluated in simulation with the realistic noise model.

The results in figure 9.22 show that both the tripod gait with rule 3 disabled and the metachronal gait are substantially less likely to achieve stable synchrony than the standard tripod gait configuration. Without rule 3, the tripod gait is less likely to be achieved at high coupling. By contrast, higher coupling is needed to obtain a stable metachronal gait.

9.7 Conclusion

This chapter has developed a distributed gait generation algorithm for a hexapod robot, as well as a physical simulation developed to assist in designing and testing this algorithm. The simulation aims to give a relatively realistic model of the real robot, including matching physical parameters such as motor torque, rigid body inertia and mass configuration, and including a simple model of sensor noise. No simulation will perfectly match the physics of a real system, and there are some aspects which proved challenging to model precisely – for example we use a relatively simple

model of the motor behaviour made available by the physics engine, and modelling certain details, such as the servo mechanics (including for example a precise model of gear friction effects) is something that could still be improved. However, the simulation model gives qualitatively similar results to the real robot when running the gait generation algorithm.

The gait generation algorithm itself consists of a set of independent, identical oscillation generators (one for each of the six legs), and a set of internal coupling rules inspired by the Walknet model. The key features of this are the independence of the oscillation generators as dynamical systems – the gait results from the emergent synchrony between the set of the oscillators – and the designed-in ability to use mechanical coupling between the legs to partially guide this synchronisation process. This aspect is most dramatically demonstrated by the ability of the system to generate a tripod gait even when all internal contralateral coupling is removed. The fact that this ceases to work when the robot is taken off the ground (thus removing the causal pathway connecting the legs mechanically), demonstrates the significance of the mechanical coupling in this case. It also suggests that mechanical coupling may have an effect on the system even in cases where we have full coupling.

In the next chapter we will see how this mechanical coupling influences the information dynamics of the system, as seen in the structure of transfer entropy statistics between pairs of legs.

Chapter 10

Information transfer in hexapod gaits

This chapter studies the causal influences acting between the components of the distributed gait generation system introduced in the previous chapter. With a more complex robotic platform, we are in a position to investigate how the techniques developed in earlier chapters apply to a system for which we do not have simple and precise causal models as in chapters 4 and 6, and which is more of a complete robotic system than the minimal examples in chapters 3 and 8.

First of all, this chapter investigates how well the expectations derived from the previous work apply in this more "realistic" scenario with regards to the relationship between transfer entropy and causation. The phenomenon of hidden information transfer (see chapters 6 and 8) is also demonstrated in a realistic system. The complexity of the current system makes the results harder to analyse than, for example, the simple swinging agent in chapter 8, however, in this new system we find an unexpected change in information dynamics when the robot is placed on and off the ground. Placing the robot on the ground introduces a causal pathway through its body morphology, but in spite of this often reduces the transfer entropy values measured between the leg angles of the robot. However, adding this external, mechanical coupling also increases transfer entropy values measured between internal variables – i.e. precisely those variables which are not directly affected by the body morphology. This results from the complex interactions of many distinct causal pathways in the system.

The robotic system introduced in the previous chapter presents in some ways a convenient target for our analysis, since the levels of causal influence can be controlled both "internally", by varying the coupling strength parameters of the gait generation algorithm, and "externally" by simply running the robot on or off the ground (thus allowing or prohibiting information transfer through the mechanical interactions of multiple legs in simultaneous contact with ground). This gives a rich variety of system types to study compared to the relatively minimal examples seen earlier in this thesis (e.g. chapters 3 and 8). However this complexity also introduces problems

of identifying the specific causal pathways along which influences are transferred. It is therefore important to first construct a principled experimental plan and set of statistical tools with which to analyse the data.

To begin with, we refine the statistical approach in section 10.1. In the second section 10.2 we discuss the ability of these techniques to demonstrate genuine causation and describe results on the simulated robot. Finally section 10.3 describes a number of ways in which the morphology and internal coupling influences the observed information dynamics, including the presence of hidden information – this analysis is extended to the real robot in section 10.3.1.

10.1 Statistical inference of causation

We have argued in previous chapters that transfer entropy is a well-justified statistic for inferring the presence of casual influences in the sense that where no causal influences are present, provided some reasonable conditions are met (chapter 4) transfer entropy will be zero. That is, in the absence of causation from X to Y:

$$\text{X does not cause Y} \implies TE_{X\to Y} = I(X_t; Y_{t+1}|Y^{(l)}_{t-l+1}) = 0 \tag{10.1}$$

This, as we have extensively noted, does not necessarily mean that a low value of transfer entropy can be taken to imply the absence of causal influence (which would be fallacious logic). Rather, the absence of transfer entropy should be taken as an *absence of evidence for* causation, rather than *evidence of absence of* causation.

Critically however, to apply this reasoning in a principled way, we should consider the ways in which the random nature of a system under study leads to inherent variance in the estimates of transfer entropy we obtain. When we apply a statistic to the data obtained from a finite run of the algorithm in simulation or on the real robot, what we have is not necessarily the true transfer entropy, but an uncertain estimate of the true transfer entropy.

We assess this uncertainty via two methods. Neither is perfect, but will give us a principled way to establish whether or not the transfer entropy values we obtain can be used to infer the presence of a causal influence.

10.1.1 Null hypothesis test

Our model of inference follows the *error statistical* approach (Mayo, 1996; Mayo and Cox, 2006) discussed in chapters 4 and 6. In order to justify the claim that X is a cause of Y on the basis of an estimated transfer entropy $TE_{X\to Y}$, we must show that the value of this estimate would be

unlikely if the putative causal influence were not present. Since we assume that the absence of the causal influence implies zero transfer entropy as in equation 10.1, this means that we are looking for the probability that the estimated transfer entropy $TE_{X\to Y}$ would be greater than the one we observed $TE^{(obs)}_{X\to Y}$ under the assumption that the true transfer entropy $TE^{*}_{X\to Y}$ is zero:

$$p = \Pr(TE_{X\to Y} \geq TE^{(obs)}_{X\to Y}; TE^{*}_{X\to Y} = 0) \tag{10.2}$$

The resulting probability p is the p-value associated with traditional null-hypothesis testing approaches. Typically this is calculated by finding the *sampling distribution* – the probability distribution of the statistic under the given assumption, whilst keeping ancillary but relevant factors (such as the size of the sample and marginal distributions) constant. In chapter 4, we discussed the sampling distribution for conditional mutual information on categorical data, *viz.* the χ^2 distribution. However, this analytic solution is not always applicable: for many statistics such as the k-NN mutual information estimator (Kraskov et al., 2004; Vlachos and Kugiumtzis, 2010, see also chapter 5) that we use to estimate transfer entropy in this chapter, the sampling distribution does not have an analytical formula.

Instead, we attempt to estimate the sampling distribution using Monte Carlo methods. This is related to the permutation approach used in chapter 3, following Marschinski and Kantz (2002), to calculate "effective transfer entropy". Given that we have obtained two sequences of values $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, we first obtain an estimate of $TE_{X\to Y}$ from the conditional mutual information form of TE (equation 10.1) which can be estimated using the k-NN conditional mutual information algorithm – this method was also used in chapter 5. In all the results here we used a history length of $l = 4$ for transfer entropy and the number of nearest neighbours k in the k-NN estimator is 5.

We then repeatedly choose random shufflings of the x series, x' – this random shuffling is expected to remove any statistical relationship between values in x' and y while the marginal distribution of x' is the same as that of x. We assume then that the transfer entropy values obtained from putting x' and y into the same estimation algorithm are thus drawn from the sampling distribution under the assumption that X and Y are statistically independent. This is similar to the permutation test described by Vicente et al. (2011).

Supposing that we perform k such shufflings, and denote the ith value obtained after shuffling $TE^{\prime(i)}_{X\to Y}$. The maximum likelihood estimate of p in 10.2 is

$$\hat{p} = \frac{1}{k} \# \left\{ TE^{\prime(i)}_{X\to Y} \geq TE^{(obs)}_{X\to Y} \right\} \tag{10.3}$$

Where $\#\{\cdot\}$ represents the number of instances where the inequality occurs. Since we are

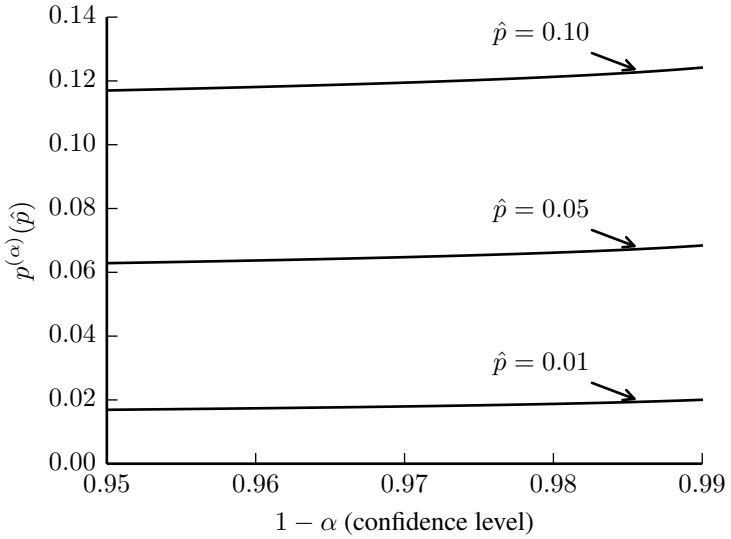


Figure 10.1: Illustration of p-value confidence interval. Given the p value estimate $\hat{p}$, we calculate an upper confidence bound on the true p value at a given confidence level $1 - \alpha$. The figure shows the change in this upper bound $p^{(\alpha)}$ for given $\hat{p}$ values obtained from 1000 random shufflings. The confidence interval is used to give a conservative estimate of the true p value – since we have used a finite number of permutations, it is possible that the true p value is higher than the estimate $\hat{p}$ obtained.

using a Monte Carlo approximation, this estimate of p is itself a random variable. To find a conservative estimate of the true p-value, we take the upper bound of the one-sided $1 - \alpha$ confidence interval for p:

$$p^{(\alpha)}(\hat{p}) = p \text{ s.t. } \mathcal{F}(\hat{p}N; N, p) = \alpha \tag{10.4}$$

Where $\mathcal{F}(k; n, \theta)$ is the binomial cumulative distribution function, i.e. the probability of k or fewer successes in n independent Bernoulli trials where the individual probability of success is θ. Thus we are finding as the upper limit for our estimate of p the value of p such that, were this the true p value, the probability of obtaining an estimate of $\hat{p}$ as low or lower than the one observed would be α. In general this upper bound, $p^{(\alpha)}$ will therefore increase as we decrease α (see figure 10.1), and for small values of α, $p^{(\alpha)}$ will generally be greater than $\hat{p}$. We find the value from equation 10.4 by finding the root of $\mathcal{F}(\hat{p}N; N, p) - \alpha$ via a binary search in the interval $p \in [0, 1]$ (note that this is guaranteed to work since $\mathcal{F} - \alpha$ is monotonically increasing with respect to p, strictly negative for $p = 0$ and strictly positive for $p = 1$, provided α is in the interval $(0, 1)$, with the exception of $\hat{p} = 1$, in which case $\mathcal{F} - \alpha = 1 - \alpha$ everywhere, in this case we let $p^{(\alpha)} = 1$).

The result of this is that we obtain a value $p^{(\alpha)}$ which we expect to be greater than the true

p value, i.e. greater than probability of a transfer entropy value as great or greater than the one observed, under the hypothesis that there is no causal influence. In the null hypothesis testing framework, $p^{(\alpha)}$ is simply a conservative estimate of the p value (conservative in the sense that it is generally larger than the true p value and thus we are less likely to obtain a spurious "false positive").

10.1.2 Bootstrapped error estimates

The second approach we take is to estimate the variability of the transfer entropy statistic itself. One way to do this is to find confidence intervals for the transfer entropy result. However, exact confidence intervals require knowledge of the sampling distribution, which we do not have. In previous chapters, where we have conducted experiments on simple, simulated systems, the variability of the estimates was indicated by the variability of the results (e.g. percentile intervals) over multiple runs of the experiment. For the current scenario, this is not practical – especially in the real robot, getting substantial quantities of data is extremely time consuming. Therefore, we investigate here the possibility of obtaining estimates of variability from a single time series record using bootstrapping.

Recall that we typically record two time series data sets

$$\begin{aligned} x &= (x_1, x_2, \ldots, x_N) \\ y &= (y_1, y_2, \ldots, y_N) \end{aligned} \tag{10.5}$$

Typical bootstrap estimates are obtained after randomly resampling the data series x and y (according to a scheme which will be described shortly), finding the value of the estimate for each resampling, and taking a percentile interval (in this case we will take the 2.5th-97.5th percentile interval) from the resulting set of estimates. It should be noted that this technique does *not* in general produce confidence intervals. There are alternative bootstrap techniques such as the BCa algorithm (DiCiccio and Efron, 1996) which purport to find the confidence intervals more precisely, however in initial tests these performed poorly on transfer entropy estimates – in general the algorithm did not converge to a stable result. This appears to result from the way BCa estimates properties of the original estimator (i.e. the k-NN conditional mutual information algorithm in this case), such as its bias, from the bootstrap distribution. This requires that the bootstrapping process itself does not introduce additional bias, however in this case it does, as we will see shortly.

We must also consider the process used to perform the resampling needed for bootstrapping. The typical approach is simply to resample the original data by drawing at random with replacement from the original data points. That is, define another series $k = (k_1, k_2, \ldots, k_N)$ where each k_i is a randomly chosen integer between 1 and N inclusive, and define the resampled data series:

$$\begin{aligned} x' &= (x_{k_1}, x_{k_2}, \ldots, x_{k_N}) \\ y' &= (y_{k_1}, y_{k_2}, \ldots, y_{k_N}) \end{aligned} \tag{10.6}$$

We would then calculate transfer entropy on x' and y'. The problem here is that we are studying time series wherein the temporal correlations between data points are significant, and would be destroyed by this resampling scheme. There are a number of time series resampling techniques that aim to avoid this problem, for example the stationary bootstrap (Politis and Romano, 1994). In these schemes, small contiguous blocks of the original time series are kept intact, and the resampled data sets are obtained by combining multiple smaller blocks to re-create a series as long as the original data sample. This aims to preserve at least some of the temporal structure of the original data, whilst producing a somewhat randomised reordering.

However, we can arrive at a somewhat simpler resampling scheme by considering the way in which transfer entropy is calculated. Recall that we are obtaining estimates of transfer entropy by way of the conditional mutual information formula in equation 10.1. That is, we rearrange the original data sets x and y in accordance with some embedding dimension l, neglecting the end points for which we do not have enough data:

$$\begin{aligned} x^t &= (x_l, x_{l+1}, \ldots, x_{N-1}) \\ y^{t+1} &= (y_{l+1}, y_{l+2}, \ldots, y_N) \\ y^{t-l+1,t} &= (y_1^{(l)}, y_2^{(l)}, \ldots, y_{N-l}^{(l)}) \end{aligned} \tag{10.7}$$

Where $y_i^{(l)} = (y_i, y_{i+1}, \ldots, y_{i+l-1})$. We then treat the data sets as ordinary samples from equivalent random variables $X^t, Y^{t+1}, Y^{t-l+1,t}$ and calculate transfer entropy via the k-NN conditional mutual information:

$$TE_{X \to Y} = I_{kNN}(X^t; Y^{t+1} | Y^{t-l+1,t}) \tag{10.8}$$

We can now resample the x^t, y^{t+1} and $y^{t-l+1,t}$ data series simultaneously in accordance with the simple sampling-with-replacement scheme. In doing so, we preserve the information structure between the three variables $X^t, Y^{t+1}, Y^{t-l+1,t}$, which by construction incorporate all the information relevant for the calculation of transfer entropy. Call the resampled data series x'^t, y'^{t+1} and $y'^{t-l+1,t}$, from which we obtain bootstrap estimates $TE'_{X \to Y}$. Note that the same resampling is applied across the three data sets concurrently, so for example, if the ith element of the resampled series x'^t is x_j (the jth element of the *original* x series), then the ith element of y'^t is necessarily y_{j+1} – this is how the temporal relationships are maintained.

There is still a problem with doing this, which is that the resampling process substantially biases the mutual information estimator. This is because, inevitably, some of the original data

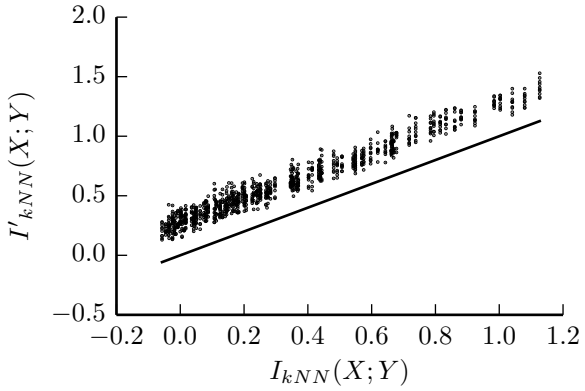


Figure 10.2: Resampling bias in mutual information bootstrap. We generate bivariate normal samples of random variable X and Y with marginal variances of 1 and various correlation coefficients ρ, obtaining 1000 data points for each sample. Using the kNN mutual information estimator (Kraskov et al., 2004), we plot the mutual information values obtained from several bootstrap resamplings $I'_{kNN}(X;Y)$ against the estimate obtained directly from the data sample $I_{kNN}(X;Y)$. The points lie above the line $I'_{kNN}(X;Y) = I_{kNN}(X;Y)$ (shown in black) indicating a strong positive bias is introduced by the resampling process.

points are repeated in the resampled series, which generally serves to increase the estimate of mutual information (as it implies a stronger relationship between the variables for those points which are repeated). This is demonstrated in figure 10.2 – note that resampling generally increases the mutual information estimate compared to the directly calculated result.

We will estimate the resampling bias as the difference between the sample mean of the bootstrap results and the original transfer entropy estimate from the non-resampled data set. We can then subtract this bias from all the bootstrap results before calculating the percentile interval. This gives a bootstrap percentile interval which encompasses the original (non-resampled) transfer entropy estimate.

This helps to deal with the bias introduced by the resampling process, but does not remove the bias of the estimator itself. The estimator bias is often downwards, as can be seen in cases where the estimator produces negative values for weak correlations (mutual information values cannot be negative, and so negative values appear to be erroneously negative). Note that negative values are obtained for mutual information for example in figure 10.2. To offset this bias, we again follow the effective transfer entropy approach (Marschinski and Kantz, 2002) and randomly permute only the x series of the original data. This, as discussed above, should enforce a true transfer entropy of zero, and thus the transfer entropy value obtained after permuting x can be taken as an estimate of the algorithm's bias and subtracted from both the transfer entropy value

and the percentile intervals.

10.2 Detecting causation

We now apply the above statistical tools to data collected in simulation. In these experiments, the control algorithm will be configured in one of the following four states (recalling the definitions of parameters and coupling rules from the previous chapter):

A. No coupling – All coupling rules are disabled. Rule 1 is inhibited by bypassing the code which prevents legs from transitioning to swing mode while the leg in front is swinging. Rules 2 and 3 are inhibited by setting the coupling constant $k_c = 0$ (note that the excitation variables will thus still be affected by the positions of ipsilateral coxa joints, but these changes will have no impact on the control commands).

B. Ipsilateral coupling only – the coupling constant k_c is set to 3 and rule 1 is enabled, however rule 3 is disabled, this time by removing the input currents that are added to opposite legs on AEP. That is, step 2 of the algorithm in table 9.1 is not applied. This means that ζ_i will be unaffected by the state of the leg opposite to leg i.

C. Full coupling. In addition to the ipsilateral coupling scenario, rule 3 is enabled as normal, the coupling k_c is still set to 3.

D. Weak coupling. All rules are enabled, but k_c is reduced to 2 to give weaker coupling strength for rules 2 and 3.

The remaining parameters are set as defined in table 10.1. Note that the swing/stance ratio $\rho = 1$ for tripod gait in all cases. The four scenarios are designed to illustrate a variety of internal coupling scenarios – it is expected that transfer entropy results should bear some resemblance to these causal structures, at least that transfer entropy should not be positive in those cases where no causal influence is present.

We obtain transfer entropy values using the kNN mutual information estimator according to equation 10.8. The system is time-lagged at one second intervals (i.e. we form time series by taking each 50th sample of the original data, since the simulation step is 20ms). We discard 100s of data at the beginning of each time series so as to obtain the stable long-term statistics, and collect 1900s of further data to use in transfer entropy calculations. A history length $l = 4$ is used, which has heuristically served as an appropriate value in the previous chapters of this thesis.

To begin with, we examine transfer entropy between all combinations of the measured leg angles α_i in the four coupling scenarios, but with the robot placed on a cuboid stand such that its

Parameter	Value
k_p	0.04
k_i	0.005
ω_{sw}	2
ω_{st}	-2
k_τ	0.9

Table 10.1: Parameter values for transfer entropy experiments in simulation.

legs move freely and do not contact the ground. This ensures that only the causal influences defined internally are present in the system. The full set of results incorporating bootstrap percentile intervals and permutation test results are shown in figure 10.3.

As discussed above, transfer entropy indicates a presence of a causal influence when the permutation test results in a small value of the upper bound of the p value. In scenario A, where there is no coupling, we expect that transfer entropy should not indicate the presence of a causal connection. There are, however, five results in scenario A which are significant at $p^{(0.05)} < 0.01$, shown in figure 10.3 (top). These are clearly false positives if taken as an indicator for the presence of a causal link, however, the stated p value has not been adjusted to take account of the fact that 30 tests have been performed just for this scenario. With so many tests, it is not unexpected that this moderate threshold would be passed by chance. At the more stringent significance level, $p^{(0.05)} < 0.001$, we find only two of these results remain significant. These still represent false positives, and suggests an even more stringent requirement would be needed to eliminate all false positives. This presents a practical problem, as the permutation test requires many more permutations to be calculated the more stringent we make the requirement – specifically, at the current 3000 permutations, the minimum $p^{(0.05)}$ value that can be obtained, occurring when none of the permutations give a greater value than the original result, is 0.0009981 – i.e. only just below the threshold of 0.001. Setting a threshold of 0.0001 for example would mean that none of the results could be confirmed as significant. Greater numbers of permutations would improve these results but are computationally expensive and for the moment we have reported the results as originally calculated.

In conditions B, C and D, at $p^{(0.05)} < 0.001$, the transfer entropy values are generally consistent with expectations – for condition B, only (and all) ipsilateral influences are detected, whereas for conditions C and D, most directions show significant causal influence in agreement with our prior knowledge that the robot is fully connected in these scenarios.

However we do have a number of "false negatives" – transfer entropy values in conditions C and D which do not strongly confirm the presence of a causal influence (at the stringent $p^{(0.05)} < 0.001$ level). This is, however, consistent with the expectation that causal influences cannot be *ruled out* merely by the absence of significant transfer entropy. We also note that moving the threshold up to $p^{(0.05)} < 0.01$ results in only one false negative for each of conditions C and D, though this does of course, as already discussed, introduce more false positives in condition A.

We now investigate what happens when the support keeping the robot's feet from touching the ground is removed. Again the four scenarios, A, B, C and D are tested, but since the robot is in contact with the ground, then there are physical influences between all legs. This would suggest causal influences in all directions, however the results in figure 10.4 show a complex structure of both significant and non-significant transfer entropy results.

Figures 10.3 and 10.4 are not ideal for examining the relationship between the transfer entropy results and the physical structure of the robot. Figures 10.5-10.8 therefore show the same results plotted on a representation of the robot's physical shape, contrasting directly the cases where the robot is on or off of the ground in each of the coupling scenarios A-D.

Thus in scenario A, where there is no internal coupling, we can see (apart from the two false positives) no transfer entropy between the legs when the robot is off the ground, but a number of significant influences when the robot is on the ground (figure 10.5).

Condition B (figure 10.6) where only ipsilateral internal coupling rules (rules 1 and 2) are applied achieves a stable gait only when the robot is on the ground (contrast figures 10.6c and 10.6f). As expected, when off the ground, transfer entropy detects causal influence only between ipsilateral legs. When on the ground, we see a tendency towards higher transfer entropy values along ipsilateral pathways compared to contralateral routes. However, looking closely at figures 10.6a and 10.6d, notice that among the ipsilateral paths (those connecting left-to-left or right-to-right), a greater number of paths are detected as significant when the robot is off the ground – generally for each pair of legs, both directions are significant when the robot is off the ground, but only one direction for each pair shows significant transfer entropy when the robot is on the ground.

Moving to condition C, where we have the full set of coupling relationships (figure 10.7, we again see a clear distinction between transfer entropy results when the robot is on or off the ground, in spite of the fact that in this case the tripod gait is successfully generated in either case. Note also that there is not a substantial difference in the *number* of pathways showing significant transfer entropy between the on and off ground conditions, but there is a clear difference in the *magnitude* of the transfer entropy values obtained. Perhaps counter-intuitively, the ipsilateral couplings especially show higher transfer entropy when the robot is off the ground – in spite of the

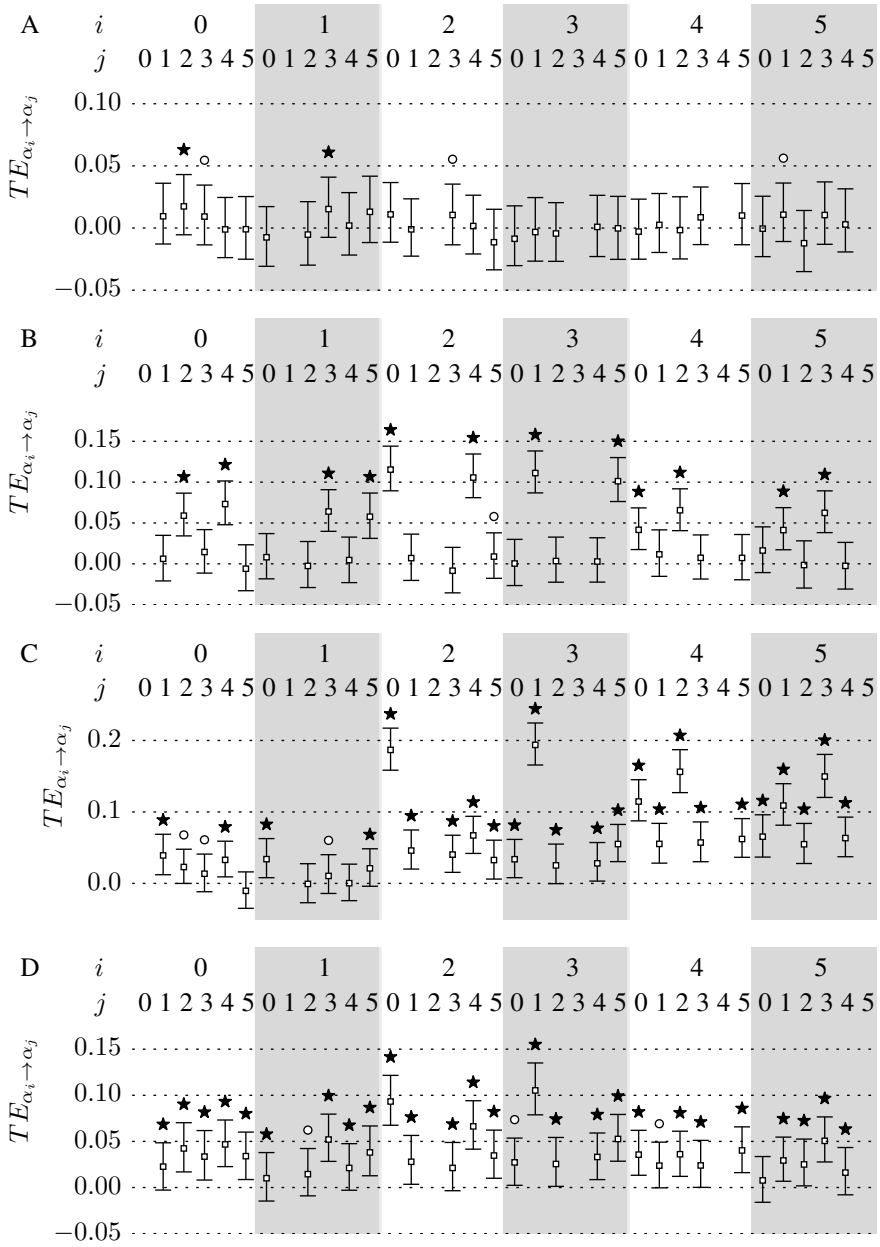


Figure 10.3: Transfer entropy between leg angles – robot not in contact with ground. Estimates of $TE_{\alpha_i \to \alpha_j}$ for all $i, j \in \{0, 1, \ldots, 5\}, i \neq j$ are calculated from a 2000s simulation of the robot under four conditions, A, B, C and D which are described in the text. The first 100s is discarded to allow the system to reach a stable dynamic, and the time series are sampled every 1s before being passed to the transfer entropy estimator. Squares and error bars show the estimate and the 2.5th-97.5th bootstrap percentile intervals. Circles represent $p^{(0.05)} < 0.01$ and stars represent $p^{(0.05)} < 0.001$ according to the permutation tests described in the text.

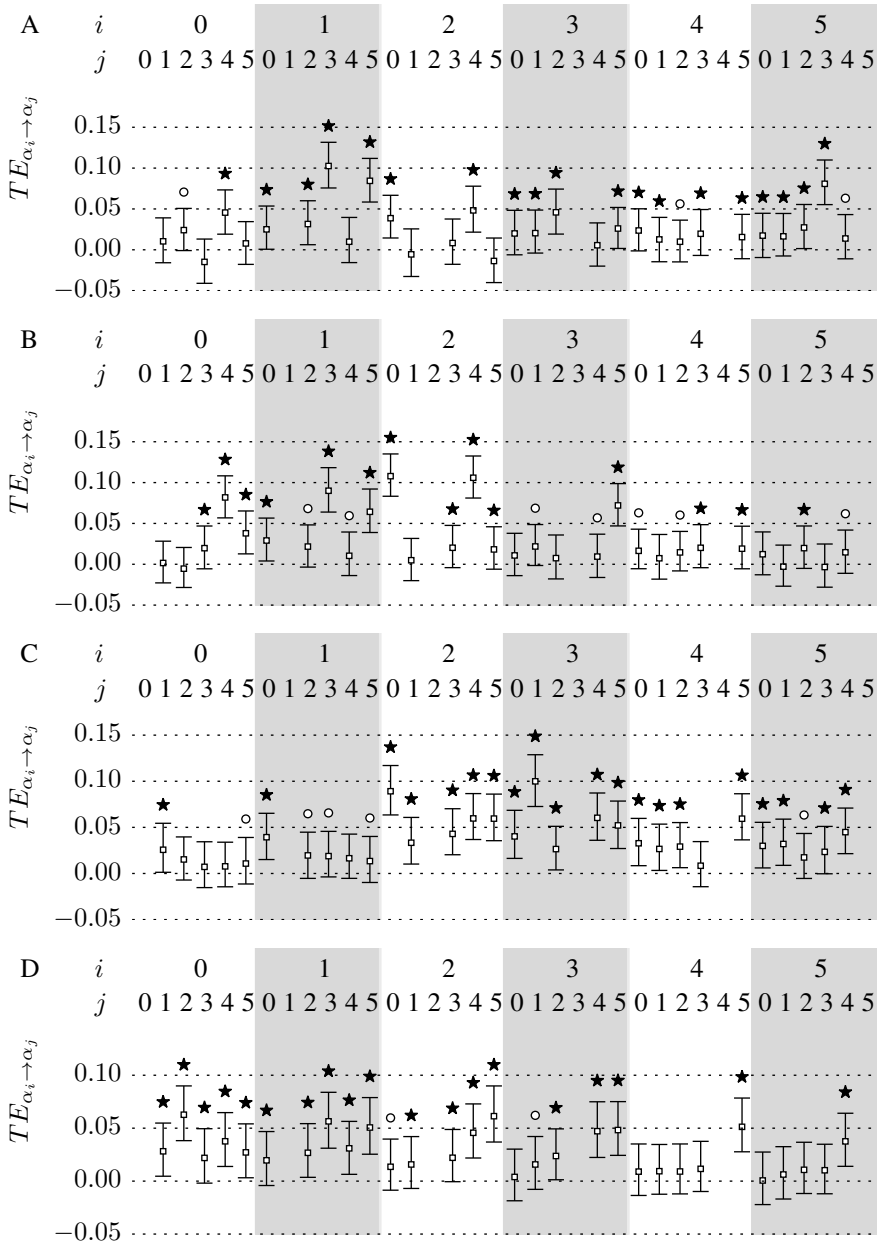


Figure 10.4: Transfer entropy between leg angles – robot in contact with ground. The same analysis is shown as in figure 10.3, but this time the support keeping the robot off the ground is removed so that the legs mechanically interact as normal.

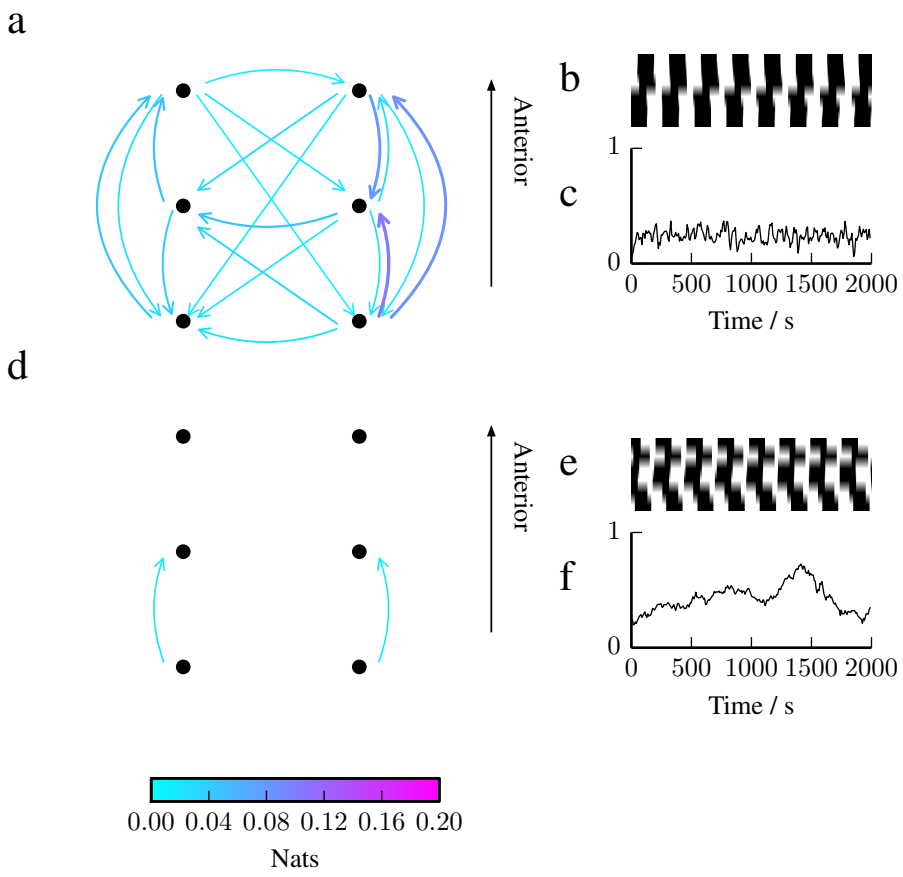


Figure 10.5: Transfer entropy summary – condition A (no coupling). **a-c**: Robot on ground. **d-f**: Robot off ground (resting on a support block). **a** and **d**: Transfer entropy on a representation of the robot's six legs, from a top down perspective with the anterior in the direction shown. An arrow is shown between leg i and j if $TE_{\alpha_i \to \alpha_j}$ was significant at $p^{(0.05)} < 0.001$ according to the permutation test. Thickness and grey-scale colour of the arrow represents the nominal transfer entropy value obtained by applying the k-NN estimator to the original data set. **b** and **e**: Representative short sections of gait as a raster plot (as in figure 9.12, black represents swing, from bottom to top legs are left-back, left-middle, left-front, right-back, right-middle, right-front). **c** and **f**: The synchrony measure σ over the complete time course, where 1 represents a successful tripod gait (see previous chapter).

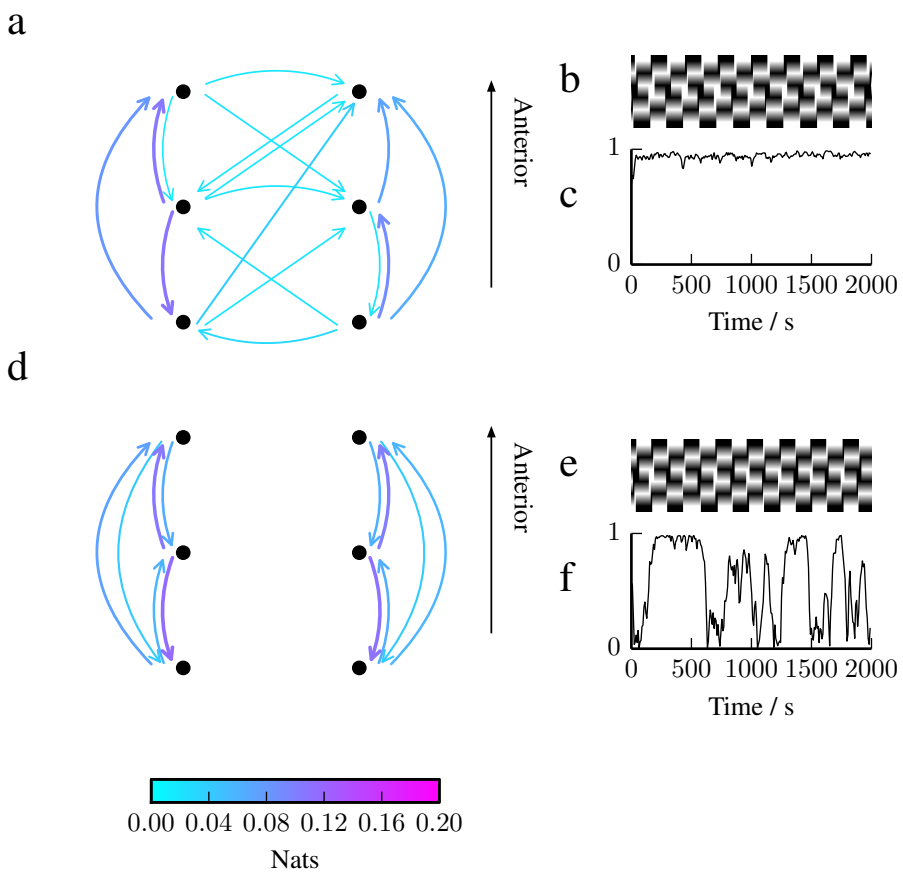


Figure 10.6: Transfer entropy summary – condition B (ipsilateral coupling only)

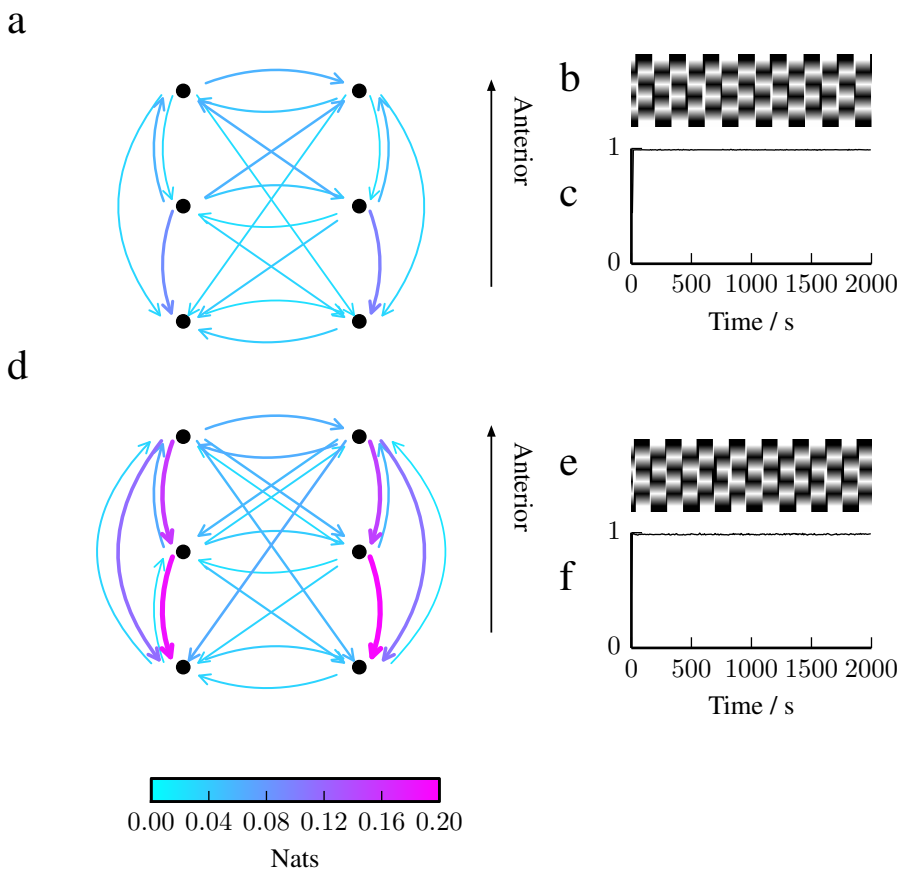


Figure 10.7: Transfer entropy summary – condition C (full coupling)

fact that there are in a physical sense fewer causal pathways in this case (when the robot in on the ground, influences can be transmitted either mechanically or through the internal input currents, off the ground only internal signals are available).

Condition D shows the case of weaker internal coupling (figure 10.8). Transfer entropy values obtained off the ground are lower than the equivalents for condition C. Another notable feature is that the *direction* of the coupling between ipsilateral middle and back legs has reversed – it is middle to back in condition C, but back to middle in condition D.

Thus, although we see many structured relationships between transfer entropy and the physical and internal coupling schema involved, much of the structure of the transfer entropy results obeys no obvious relation to the manner of the underlying coupling. Aside from the absence of transfer entropy that is generally obtained in cases where absolutely no causal influence is present, the magnitude and statistical significance of the transfer entropy results does not predictably correspond to the known facts of the system. For example, there are multiple cases where significant transfer entropy disappears when the robot is placed on the ground rather than kept in the air, even though doing so can only add (and not take away) causal pathways (since the internal causal

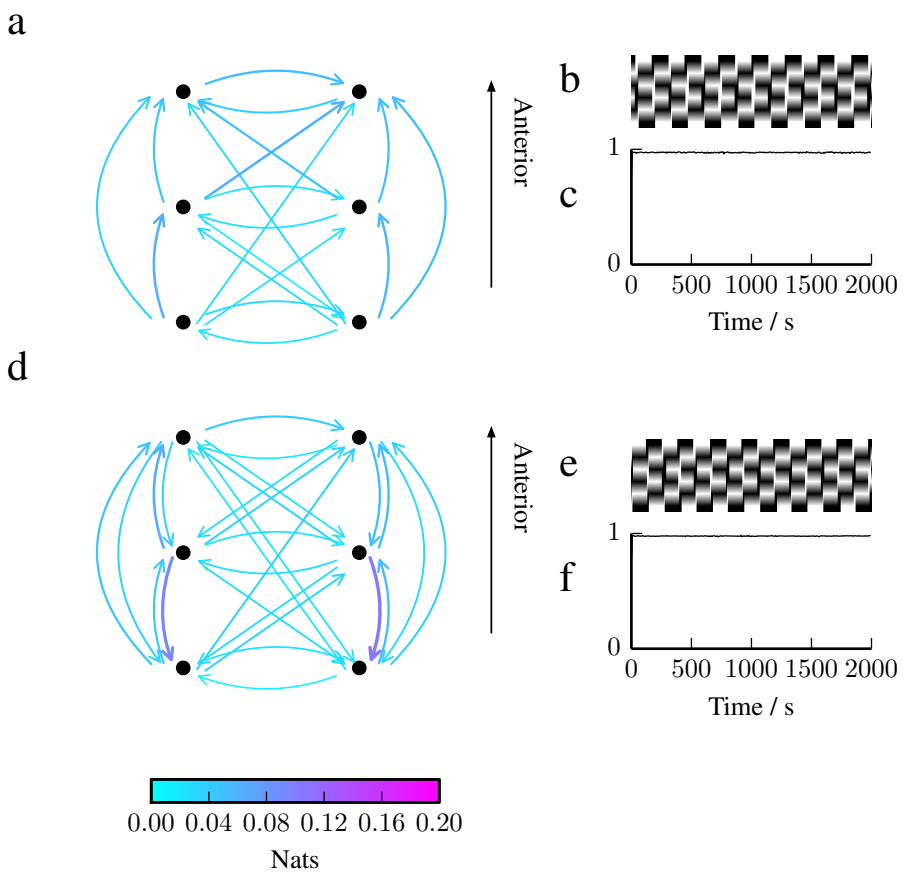


Figure 10.8: Transfer entropy summary – condition D (weak coupling)

pathways are kept constant). The reversal of the medial-posterior transfer entropy results between conditions C and D when the robot is on the ground suggest that changing the internal coupling strength (recall that condition C has $k_c = 3$ whereas D has $k_c = 2$) can both add an remove significant transfer entropy results.

10.3 Hidden information transfer

In previous chapters we discussed a phenomenon of "hidden information transfer" (chapters 6 and 8). We will test for the presence of this type of structure in the current system.

The phenomenon we are calling hidden information transfer takes the form of a chain of causal influences, such as $A \rightarrow B \rightarrow C$, where the information transfer at the more extreme points (e.g. the transfer entropy from A to C) is greater than that at intermediate points (e.g. from A to B). Previously, we have hypothesised that this is likely to occur in cases of strong synchrony, where nodes that are close together are more likely to synchronise more perfectly, leading to reduced detectability of the causal influence. The nodes that are further apart may be synchronised, but more weakly so, and thus relationship between them is more complex and easier to detect.

In the present scenario, the situation is complicated by the fact that there are in general multiple causal pathways between any two points. This is especially true when the robot is in contact with the ground and the mechanical contacts, in principle, permit information flow between the legs, bypassing the internal coupling rules. We will return to this point after discussing some preliminary results.

We begin by taking, in addition to our measurements of transfer entropy between the measured coxa joint angles ($TE_{\alpha_i \rightarrow \alpha_j}$), the transfer entropy from coxa joint angles to excitation variables, $TE_{\alpha_i \rightarrow \zeta_j}$. Recall that the excitation variables ζ control the internal coupling for rules 2 and 3 – without ζ, the individual legs' states only affect each other mechanically (when the robot is on the ground) and via rule 1 (for ipsilateral legs only, and provided rule 1 is in force). Without these pathways, α_i can only influence α_j by first influencing ζ_j (this can be seen from the structure of the controller described in section 9.3.2). This makes the difference $TE_{\alpha_i \rightarrow \alpha_j} - TE_{\alpha_i \rightarrow \zeta_j}$ a candidate for detecting hidden information – if the first term is higher than the second, then we have strong evidence that α_i causes α_j, yet less support for the hypothesis that α_i causes ζ_j, in spite of the fact that α_i (in at least some cases) can only influence α_j by first influencing ζ_j.

To begin with, we attempt to test for causation between leg angles and excitation variables using the same transfer entropy analysis as before. This section uses data from the same simulated trial runs as in section 10.2. Again tests are conducted both with the robot raised from the ground (figure 10.9) and in contact with the ground (figure 10.10).

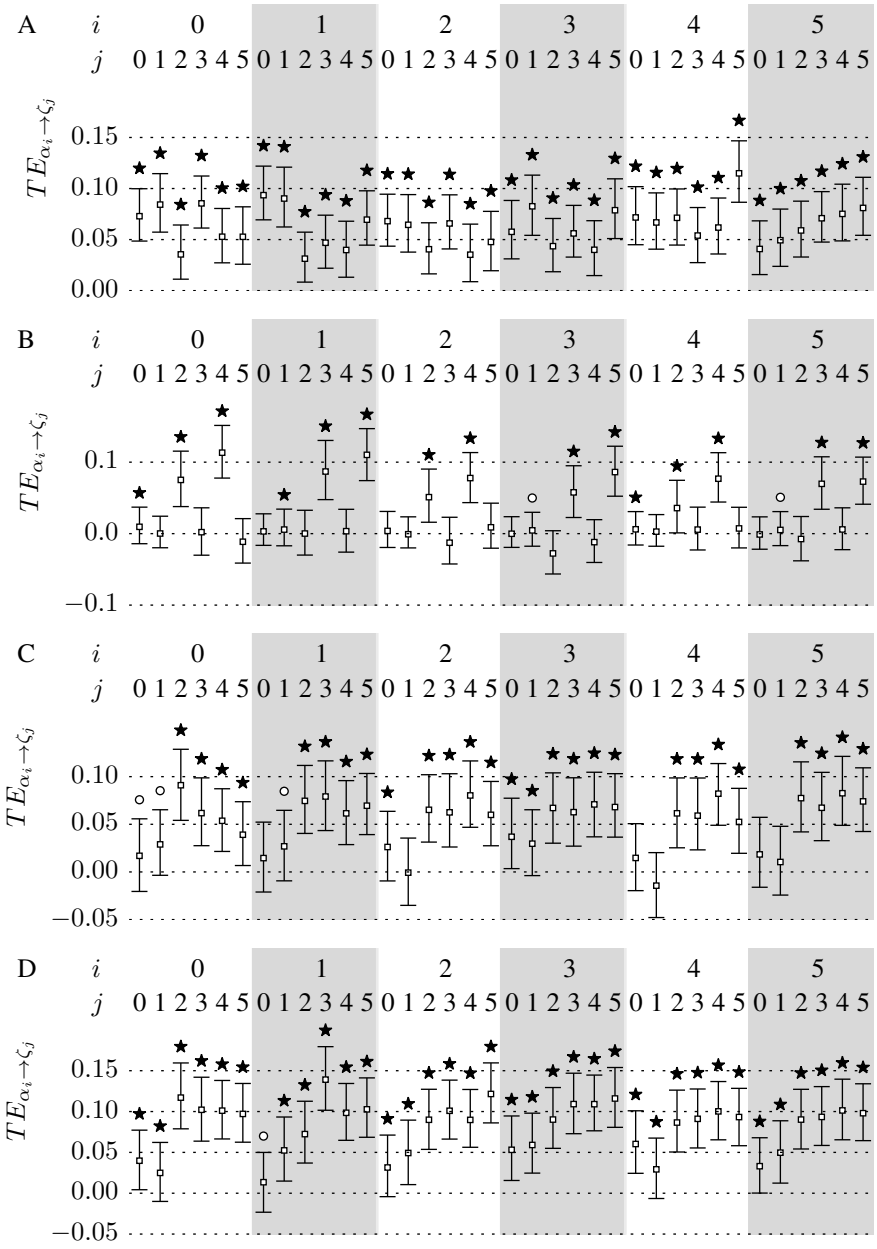


Figure 10.9: Transfer entropy between coxa angles and excitation variables – robot not in contact with ground. The results of permutation tests and bootstrapped percentile intervals, calculated as described in the text, for transfer entropy $TE_{\alpha_i \to \zeta_j}$ between pairs of legs i and j.

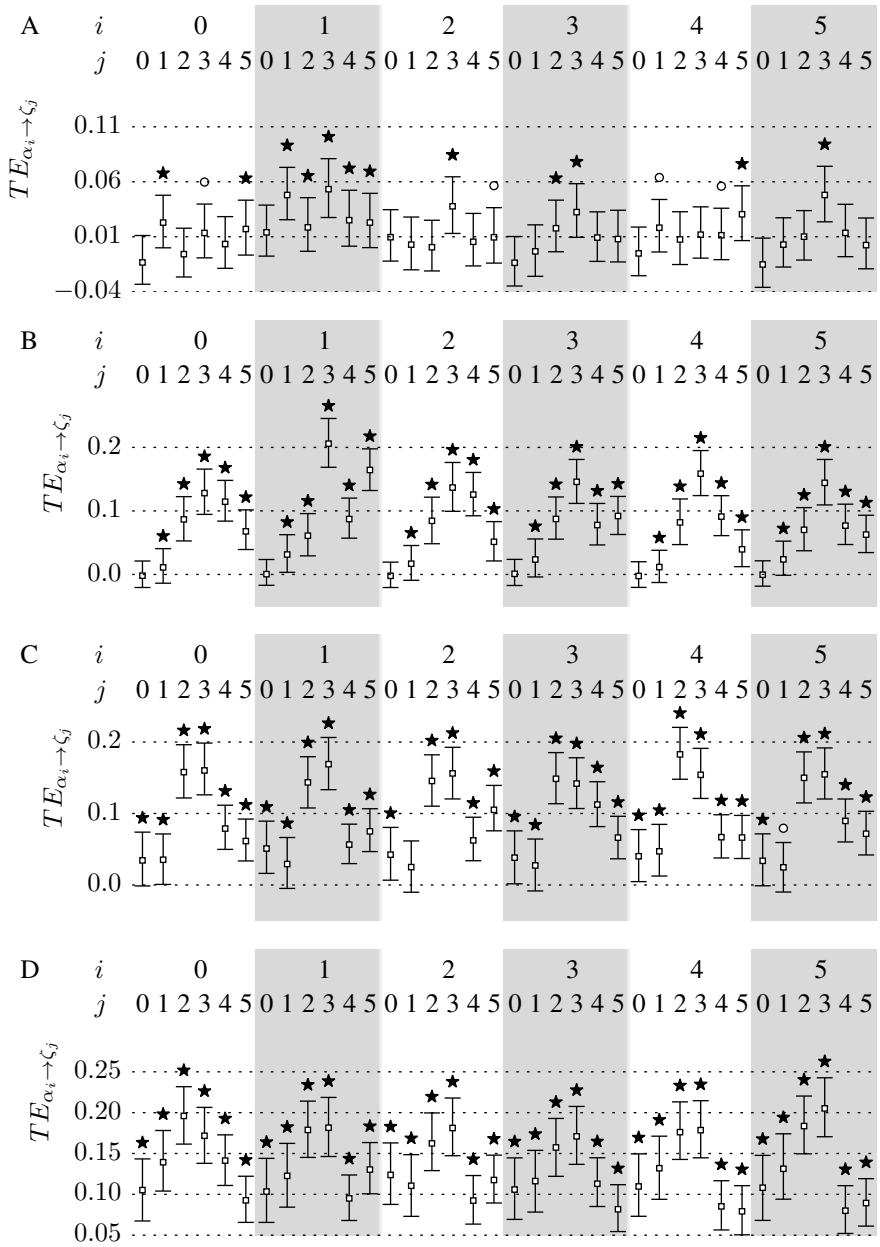


Figure 10.10: Transfer entropy between coxa angles and excitation variables – robot in contact with ground.

Recall that in scenario A, the coxa joint angles are allowed to influence the excitation variables ζ – coupling is inhibited by setting $k_c = 0$ at a later stage in the algorithm, such that the value of ζ_i multiplied by zero before being used to update the target velocity ω_i of the associated leg. Thus causal influences are in general present for scenario A in figure 10.9. However, contralateral influences are disabled in scenario B by not modifying ζ_i according to the AEPs of the opposite leg, and thus (while the robot is off the ground), we do not see contralateral influences between coxa angles and excitation variables.

Again, to get a clearer picture of the structure of these results, it is helpful to plot the transfer entropy values against a representation of the robot. Furthermore, we have hypothesised that hidden information is likely to occur under strong synchrony. Therefore, a measure of the general synchrony level between pairs of legs is useful. Unlike the global synchrony measure σ, which evaluates how well a particular gait is achieved, we are interested here in a pair-wise measure of how closely the behaviour of two legs correlate, notwithstanding the (for our purposes unimportant) phase difference between the signals. We evaluate this using the empirical cross-correlation function:

$$\alpha_i \star \alpha_j(\tau) = \sum_{t=0}^{T} \alpha_i(t)\alpha_j(t+\tau) \tag{10.9}$$

With the sum taken over a chosen set of sample times. We take $t = 0$ to represent the sample at 200s into the simulation and $t = T$ at 250s – this performs the calculation over a representative portion of the time series. Here the time series is sampled at 50Hz – i.e. there is no "lagging" (downsampling) as there is with the data used for transfer entropy calculations. There is maximum at some time difference τ, which can be taken as a measure of the similarity between α_i and α_j when the phase difference is ignored.

Figures 10.11 and 10.12 (robot off and on the ground respectively) show a comparison of this synchrony measure against the previously obtained transfer entropy values $TE_{\alpha_i \to \alpha_j}$ and $TE_{\alpha_i \to \zeta_j}$. Again, arrows are only plotted where the transfer entropy value is significant at $p^{(0.05)} < 0.001$. Hidden transfer entropy, as discussed above, is taken to be represented by the difference $TE_{\alpha_i \to \alpha_j} - TE_{\alpha_i \to \zeta_j}$. However, we do not have a direct significance test for hidden information. Instead, we consider that by hidden information we mean strong evidence for causation between the "outer" variables ($\alpha_i \to \alpha_j$) and weaker evidence among the "inner" variables ($\alpha_i \to \zeta_j$). Thus we regard the hidden information value as significant (and draw an arrow) if the result for $TE_{\alpha_i \to \alpha_j}$ was significant but the result for $TE_{\alpha_i \to \zeta_j}$ was not significant.

The results in figure 10.11 are clearest in terms of our current understanding of hidden information transfer. Note first the pair-wise synchrony measure shows the strongest overall synchronisa-

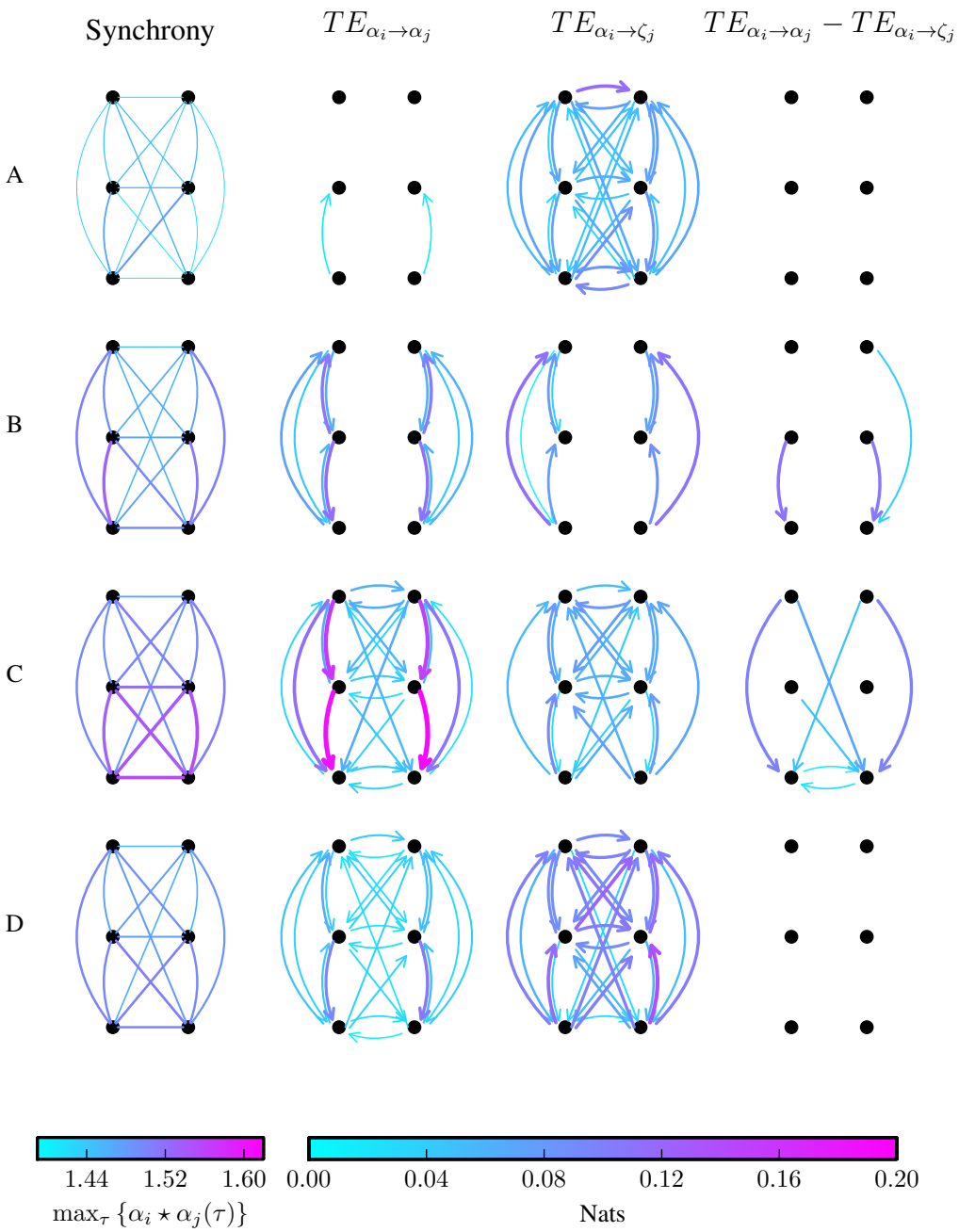


Figure 10.11: Hidden information results – robot not in contact with ground. For each of the four coupling scenarios A-D described in the text, we show the maximum pairwise synchrony measure between each leg, along with the transfer entropy between coxa angles $TE_{\alpha_i \to \alpha_j}$ and the transfer entropy from coxa angles to excitation variables $TE_{\alpha_i \to \zeta_j}$. "Hidden information" is regarded as the difference between these two variables, plotted in the right hand column. For TE results, only values pathways determined to be significant as described in the text are shown with an arrow.

tion values under the "full coupling" scenario (condition C, where all coupling rules are enabled at full strength). This is also where we see the most hidden information pathways – since in this case there is no mechanical coupling, the coxa angles can only influence each other through the internal coupling rules, and only rule 1 allows direct influence among the coxa angles α – rules 2 and 3 require than any influence of α_i on α_j "passes through" the internal excitation variable ζ_j. More than this, rule 1 does not apply contralaterally – there is no way for coxa joints on the left hand of the robot to influence coxa joints on the right hand side without making use of the excitation variables, yet we do see contralateral information transfer among the coxa joints in some of the cases where there is no significant information transfer between the coxa joints and excitation variables. In condition D, where the coupling is weakened, this phenomenon disappears – by contrast, here we have the more intuitive higher transfer entropy values going from coxa angles to excitation values, than coxa angles to coxa angles.

The situation where the robot is on the ground is radically different, as shown in figure 10.12. Here there are very few significant hidden information transfer results in the conditions where a successful gait is generated (B, C, and D). Moreover, there is a clear pattern of high values for $TE_{\alpha_i \to \zeta_j}$ and lower values for $TE_{\alpha_i \to \alpha_j}$ in conditions B, C, and D. For condition A, with no internal coupling, we do appear to see a number of apparent hidden information transfer results, however these are less meaningful than before – recall that in condition A we do allow α_i to influence ζ_j, but we do not allow ζ_j to influence the downstream α_j. Thus, in condition A, the measure we have taken does not actually correspond to our definition of hidden information transfer since α_i does not influence α_j "through" the excitation variables ζ_j.

Of course, in conditions B, C, and D we no longer have a good fit for the concept of hidden information transfer either, since there are now at least two pathways that α_i can influence α_j along – both through the internal variable ζ_j and also through unmeasured mechanical influences. However, an interesting feature of this result is that the addition of the extra causal pathway between coxa joints does not increase the information transfer between the coxa joints themselves, but it *does*, generally, increase the information transfer along the *internal* pathways (compare the third column, $TE_{\alpha_i \to \zeta_j}$ between figures 10.11 and 10.12). This was unexpected, since the internal pathways are not directly affected by changing whether the robot is in contact with the ground.

Perhaps even more surprising is the presence of contralateral information transfer to the excitation variables in condition B – here rule 3 is disabled and so there is no direct influence of coxa angles on contralateral excitation variables, however we see high, almost all significant values for $TE_{\alpha_i \to \zeta_j}$ under condition B. This is not a false positive of course, since there is an *indirect* route between a coxa angle and a contralateral excitation variable – namely, one that passes (contralat-

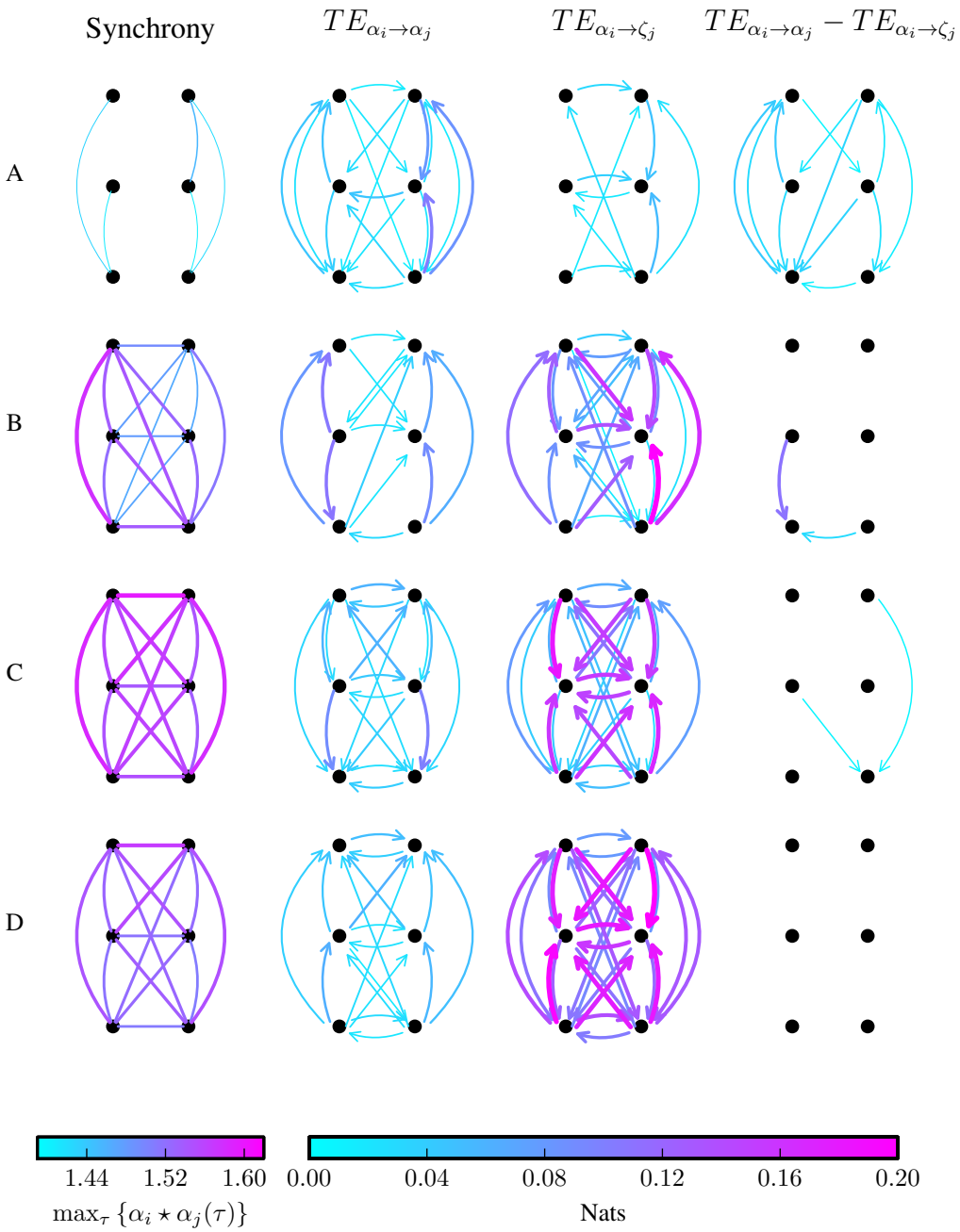


Figure 10.12: Hidden information results – robot in contact with ground. Synchrony and transfer entropy values as in figure 10.11, calculated from the results obtained when the robot is in contact with the ground.

erally) through the mechanical coupling first (i.e. directly along $TE_{\alpha_i \to \alpha_j}$), and then ipsilaterally (via either rule 1 or rule 2) to the internal excitation variable. What is notable, however, is that the contralateral values for $TE_{\alpha_i \to \alpha_j}$ in condition B are typically small, and often non-significant, and so the first step of this pathway is again not as visible in the transfer entropy results. This is suggestive of another example of hidden information transfer, in this case not directly measured. Unfortunately, it does not seem possible using the measured data we have to directly capture this hidden information transfer numerically, since there several combinations of individual influences which could facilitate the contralateral influence of some α_i on some other (on the opposite side) ζ_j.

10.3.1 Results for the real robot

We investigate whether similar results are obtained from the real robot as in the simulation. Here we look only at conditions B (half coupling), C (full coupling) and D (all coupling rules enabled, but with weaker strength than C). Condition A has not yet been tested with the real robot. Again we aim to test the robot generating a tripod gait. However, for successful gait generation, the real robot requires slightly different parameter settings to the simulation in order to achieve stable gaits – as was described in section 9.5.1 when the gait algorithm was originally tested on the real robot Here, the parameters given for the simulation in table 10.1, are the same for the real robot except for the following (further discussion of the reasons for these changes is given in section 9.5.1):

- k_i – The integral control parameter is changed from 0.005 in simulation to 0.01 in the real robot, generally reducing the controller error.
- k_τ – The decay constant for the excitation variables is increased from 0.9 in simulation to 0.95 in the real robot, meaning the excitation variables decay more slowly.
- k_c – The coupling parameter was set to 3 for conditions B and C and 2 for condition D (the "weaker coupling" scenario) in simulation. In the real robot, we use 1.5 for conditions B and C and 1 for D.

Due to these changes in parameters, as well as the simple fact that the simulation is not a perfect reconstruction of reality, we naturally expect some deviation between simulation and real robot results. However, both the real and simulated robots produce similar tripod gaits when either all coupling rules are enabled (conditions C and D) and also in condition B (half coupling) provided the robot is on the ground.

Since it is naturally more time consuming to collect data from the real robot than simulation, we calculate transfer entropy results from 300s of data samples at 50Hz in the real robot (as

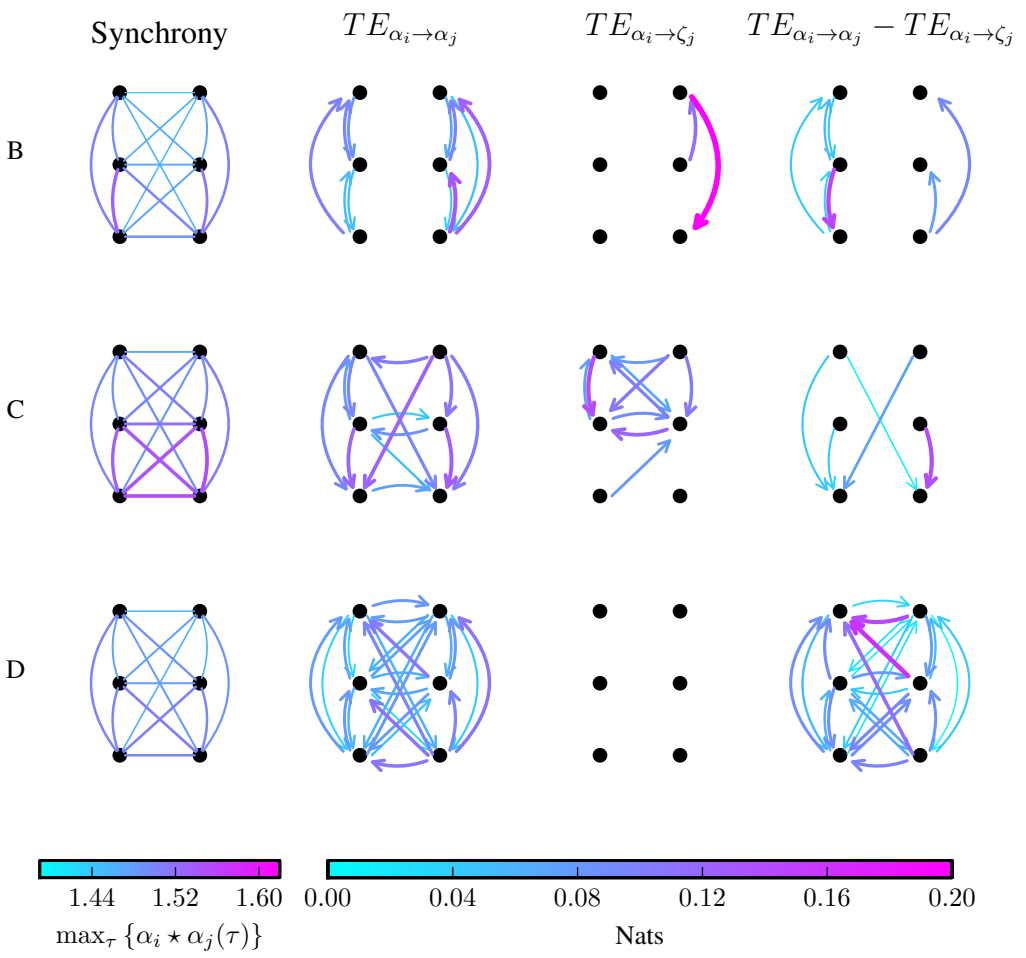


Figure 10.13: Hidden information in the real robot, not in contact with ground.

opposed to 1900s in simulation), after discarding the first 60s (rather than 100s) from the time the gait was initialising. As in the simulation, we only take every 50th sample for calculations of TE, so that we have samples 1 second apart.

The results are presented in the concise form in figures 10.13 and 10.14 – comparable to the equivalents for the simulation in figures 10.11 and 10.12. For completeness, the full set of results with bootstrap intervals are appended at the end of this chapter (figures 10.15 to 10.18 in section 10.5).

The results pertaining to the transfer entropy between joint angles, $TE_{\alpha_i \to \alpha_j}$ show some similarities between simulation and the real robot – for example with rule 3 disabled (condition B), there is still no contralateral transfer entropy when the robot is off the ground, so in the genuine absence of causal influence we find zero transfer entropy. In other cases, there are generally significant transfer entropy values across most pairs of legs.

We also see that for conditions B and D, the internal transfer entropy $TE_{\alpha_i \to \zeta_j}$ is significant along more pathways when the robot is placed on the ground (i.e. there are generally more arrows

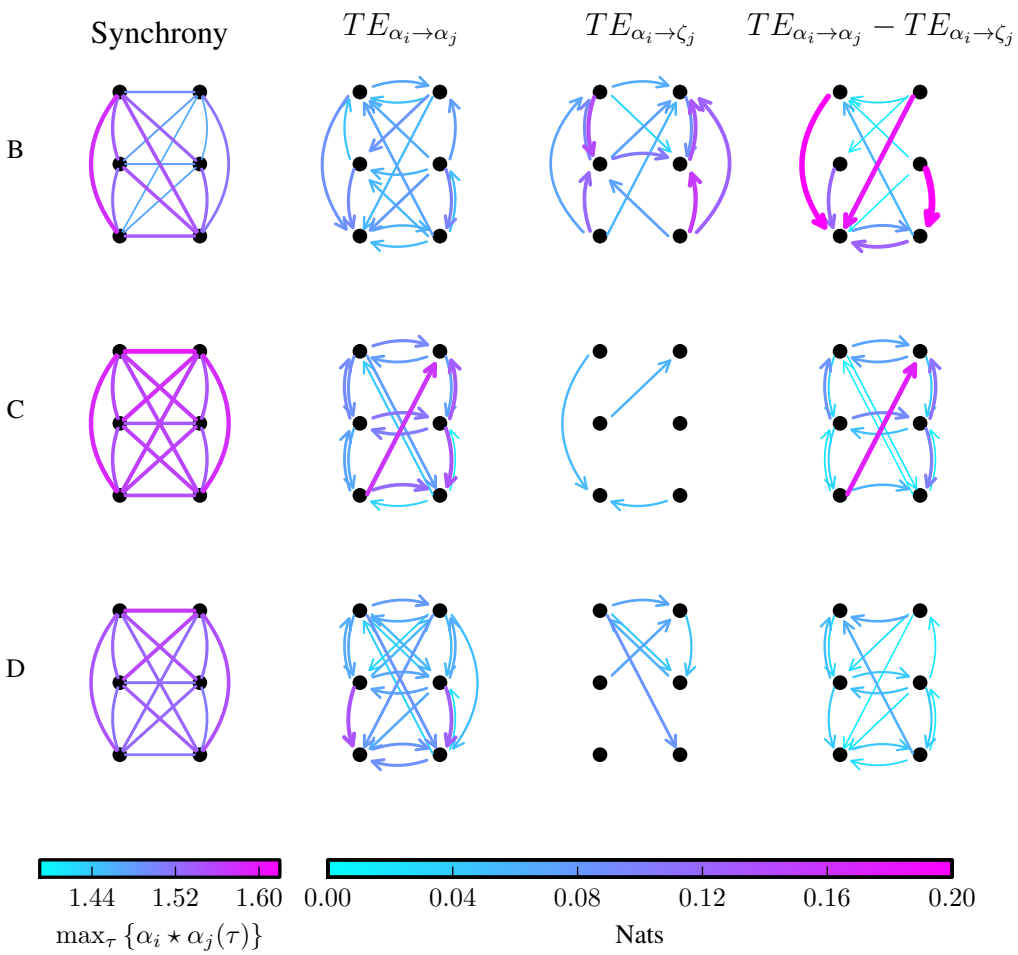


Figure 10.14: Hidden information in the real robot, in contact with ground.

in the third column of figure 10.14 than 10.13).

However, there is a substantial difference in the overall number and magnitude of the internal transfer entropy results between the real and simulation situations – although there are significant values of $TE_{\alpha_i \to \zeta_j}$ in the real robot, there are generally fewer and they are lower in the real robot than the equivalents in the simulation. This seems to be the primary reason for the larger number of hidden information results in the real robot. Recall that in simulation, $TE_{\alpha_i \to \alpha_j} - TE_{\alpha_i \to \zeta_j}$ generally did not achieve significant positive values when coupling is present and the robot is on the ground in simulation (figure 10.12), but on the real robot we have many significant hidden information results (figure 10.14).

These differences may partly be explained by the changes made to the parameters in the real robot, and due to the aspects of the real robot physics not accurately captured by the physics simulation. A possible area for exploration is the contact joints between the robot and the ground – the real robot has rubber "feet" on the end of each leg, and was run on plastic flooring tiles in the laboratory. In simulation, this is naively modelled by a simple friction model, but it is one of the aspects of the simulation that was not carefully matched to the real experiment situation. That is, the friction parameters in simulation were chosen to give a reasonably realistic appearance, but unlike say the motor torque model where the step response of the simulation was keyed to real robot results (section 9.2.1), no such validation step was performed for the floor friction model. It is known that ground contact effects can have a significant impact on transfer entropy results (Schmidt et al., 2012) – indeed this could be seen as one of the features of morphological computation illustrated by the transfer entropy approach.

10.4 Conclusion

This chapter presented a well justified set of tools for inferring the presence of causation from transfer entropy results. In contrast to the work in earlier chapters, extra focus has been placed on determining the statistical significance of transfer entropy results, rather than only looking at the *prima facie* value of transfer entropy obtained. This has enabled us to look, in a more principled way and in more detail, at the structure of information transfer in a system that contains a much greater level of structural complexity than the minimal examples presented in earlier chapters.

We have seen the phenomenon of hidden information transfer that was predicted in accordance we the theoretical examples presented earlier in this thesis. However, this was primarily seen in the real robot, but not in the simulated robot, in spite of the behavioural similarity of the two.

Perhaps the more interesting result was more unexpected – the contrast between the information dynamics of the robot on and off the ground. In simulation, placing the robot on the ground

reduced the transfer entropy values between coxa joint angles (which are more or less directly coupled via the body morphology and ground contact) but increased in those transfer entropy measures which looked at internal variables. Information dynamics has previously been proposed as a tool for measuring *morphological computation* (chapter 7). This is, in a sense, confirmed here, but counter-intuitively, the scenario where the morphology of the robot has a greater influence (when it is on the ground), did not lead to higher information transfer between those components, namely the coxa joint angles, which have the most direct connection to the body morphology. Rather, the transfer entropy from coxa joints to internal variables was seen to increase – suggesting that the indirect causal influences are easier to detect.

This reflects the difference between the interpretations of information transfer as representing causal strength or inferential warrant: previously, we studied the changes in information transfer as coupling coefficients were changed, giving an intuitive and easily comparable change in true causal strength. Here, we have also looked at what happens when the number of direct and indirect routes for causal influences is changed.

10.5 Full real robot transfer entropy results

Here we include the full set of transfer entropy results for the real robot. These figures are generated from data in the same way as those shown in section 10.2.

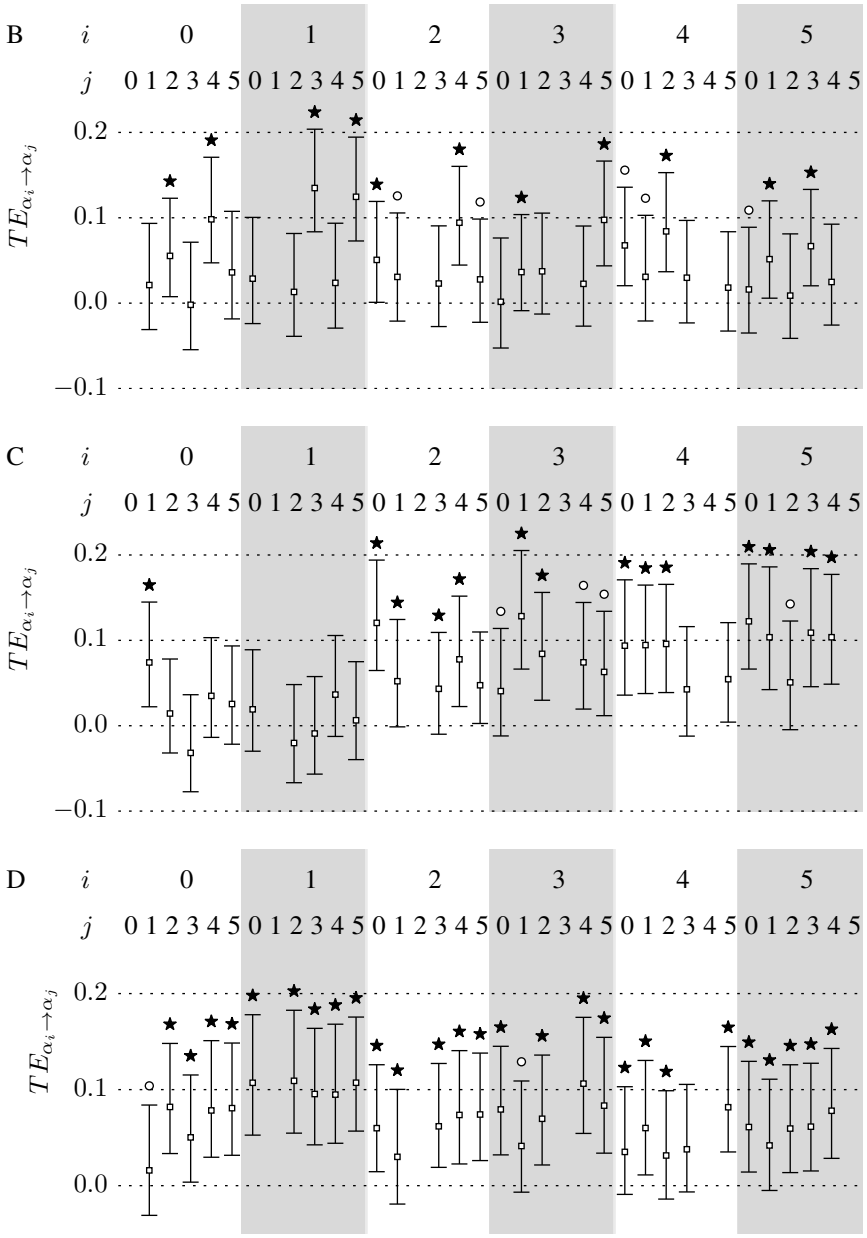


Figure 10.15: Transfer entropy between coxa angles on real robot not in contact with ground. Values are calculated in three conditions – B, C, and D as described in section 10.2. This and all figures in this section show transfer entropy results in the same manner as figure 10.3 – specifically, central squares show the *prima facie* transfer entropy value, the error bars give the 2.5th-97.5th percentile bootstrap interval and statistical significance is represented by the overlying marks (circle for $p^{(0.05)} < 0.01$ and star for $p^{(0.05)} < 0.001$).

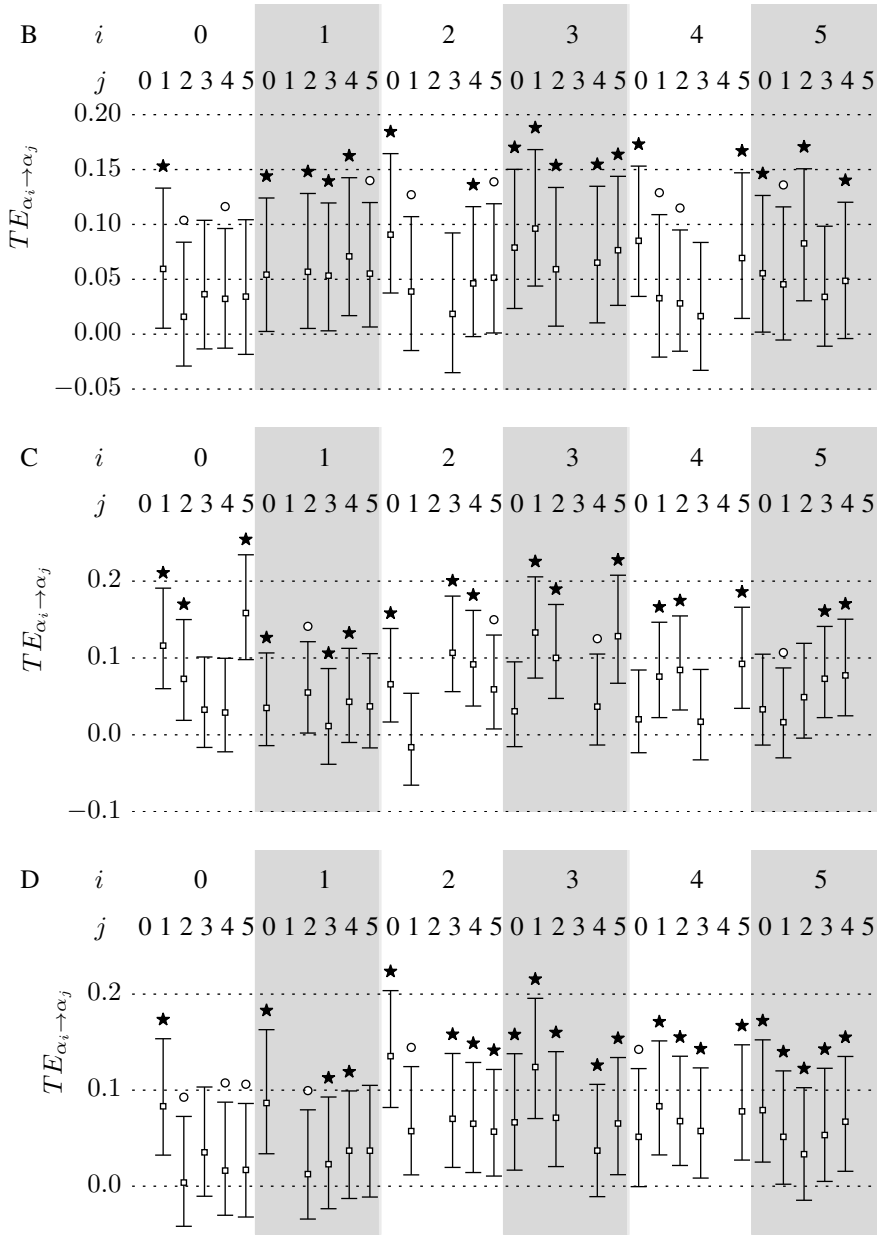


Figure 10.16: Transfer entropy between coxa angles on real robot in contact with ground.

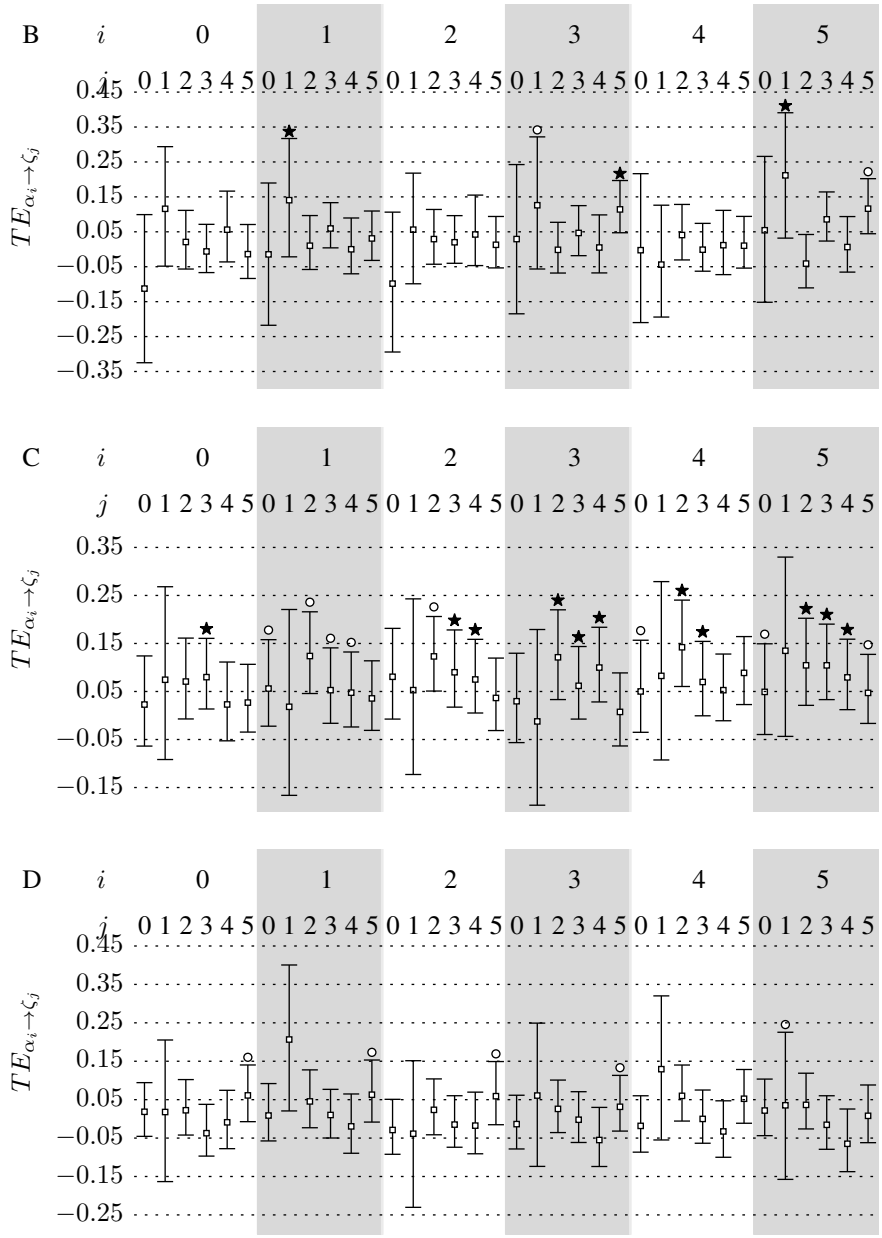


Figure 10.17: Transfer entropy between coxa angles and excitation variables on real robot not in contact with ground.

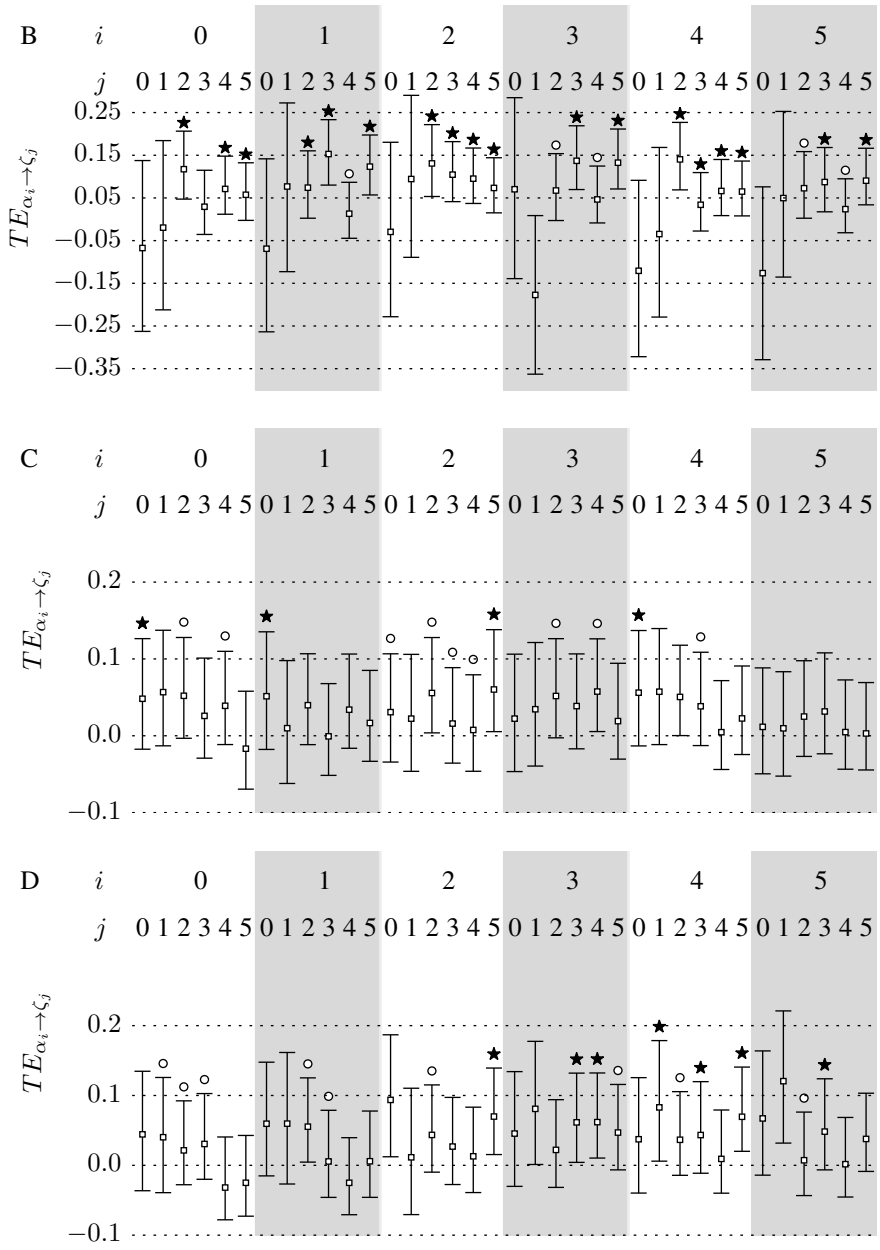


Figure 10.18: Transfer entropy between coxa angles and excitation variables on real robot in contact with ground.

Chapter 11

Discussion

This thesis has given an account of statistical and particularly information theoretic approaches to detecting causation, with a view to applications in studying autonomous robotic systems. While much of the work has tackled causation from a fairly general perspective, the focus on gait generation has led to particular emphasis on synchronisation. In the work on legged locomotion, we have demonstrated a decentralised architecture consisting of identical oscillation generators coupled in such a way as to produce symmetric patterns of synchrony similar to animal gaits. This architecture was designed such that it makes use of mechanical feedback through the body as part of the coupling between components of the body. This allows us to study the causal and information processing role of the body in generating the gait.

This chapter briefly summarises some of the main results and contributions of the thesis, then goes on to suggest some possible research problems which could make further use of the ideas presented in this thesis.

11.1 Information, complexity, and inference of causation

In chapters 4 and 6, we have tried to distinguish between the roles of information as a measure of *complexity* and as a tool for *inferring causation*. Both these viewpoints are valid, but the distinction between the two is critical.

As a measure of complexity, information tells us about how much of the uncertainty in regards to one variable can be reduced by knowing the value of another. This requires that the variables actually have some non-trivial uncertainty (more precisely, *entropy*) from the outset. Complexity in this context arises when a variable can be used to predict something *that would not otherwise be trivial to predict*.

For any causal interpretation, we have argued that a layer of causal reasoning, which can be

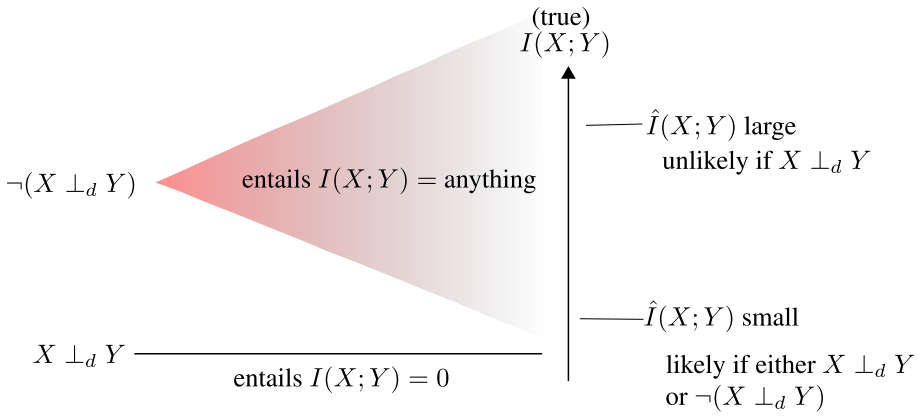


Figure 11.1: A schematic of severity-based causal inference. Working from left-to-right, a hypothesis of no causal connection between X and Y, written here as $X \perp_d Y$ entails that the mutual information $I(X;Y)$ must be zero, and a hypothesis that there does exist a causal connection on a graph – $\neg(X \perp_d Y)$ – while it *permits* a mutual information $I(X;Y)$ greater than zero, does not (by itself) provide any reason to think that the true mutual information value will take any particular value – it could still be arbitrarily close to zero. Reading right-to-left gives the inferential interpretation: if we observe a high value of empirically measured mutual information $\hat{I}(X;Y)$, we can say that the hypothesis that there is a causal influence has passed a *severe* test, because if there was not a causal influence (i.e. $X \perp_d Y$ is true), a high value of mutual information is unlikely. But, if we observe a low value $I(X;Y)$, neither the hypothesis of of causal influence, nor the hypothesis of no causal influence, has received a severe test, because either could with some probability result in a low value of mutual information. That is, low mutual information is used to mean "no inference" regarding causation, rather than to infer the absence of causation.

provided through causal Bayes nets, must be added. The formal definition of *information transfer* discussed in chapter 4 shows how this can be done. However, we argue that it is crucial to bear in mind that *not all causal relationships have to be complex*. That is, it is not unexpected for information theoretic statistics which are quite reasonable as tools to infer causality may still produce small values when a causal influence is present.

The fact that some information theoretic methods, notably transfer entropy, have a tendency to do this has been noted or at least alluded to in various parts of the literature already. The central contribution of this thesis in this regard is to put this point on a sound theoretical basis. That is, we have given an argument as to *why* information transfer can been seen as a valid statistic for causal inference in spite of this drawback, as a consequence of the *severity requirement* discussed in chapters 4 and 6. Recall that (following Mayo (1996)) we said that we can pass a hypothesis H on the basis of some data D "with severity" if the data accords with H and further, if H

were false, data like D would be unlikely to arise. If we apply this approach to inferring causal influences in graphical models, we arrive at the conclusion that low mutual information (including low conditional mutual information and transfer entropy) is best taken to give no strong inference either way as to the presence or absence of causation, but that high information may allow us pass a hypothesis of causal influence with severity (see figure 11.1 for an illustration).

A consequence of this is that what we are thinking of here as the complexity of a relationship can also be thought of as a kind of *epistemological clarity* of the relationship – causal influences which are "complex" in the sense of producing high information transfer, should be easy to detect by experiment, and also easy to rule out in the cases where they are absent. For example, the cases we have often discussed of strong synchrony are examples of strong causal connections that do not have such epistemological clarity, or equivalently, would not be seen as "complex" – yet they are indeed instances of causal influence. This places us on a substantially firmer footing for making use of such tools going forward – see below (section 11.4.1) for one suggestion of how to make use of this.

11.2 Alternative approaches to detecting causation

A major effort of this thesis has been to consolidate various techniques where possible into a similar framework. This is one area where information theory, giving a common mathematical formalism, seems to provide a particular advantage. It allows us to focus on the *structure* of the relationships between multiple variables without, by virtue of it being "model-free" being dependent on the specific functional *form* of those relationships.

In particular, we have constructed a general concept of information transfer and used it as a common framework for interpreting other proposals for detecting causation, namely *transfer entropy* (Schreiber, 2000) – the main topic of this thesis, *information flow* (Ay and Polani, 2008) and *convergent cross-mapping* (Sugihara et al., 2012).

Information flow as defined by Ay and Polani (2008) makes explicit use of causal Bayes net theory to formalise its relationship with causation. Specifically it incorporates *intervention* on a causal Bayes net, as defined by Pearl (2009), to ensure that high information flow can only occur in the presence of a causal influence (see chapter 4).

The first part of our argument is to bring information transfer into the same causal Bayes net formalism used for information flow. Information transfer is defined in cases where there is no intervention, and thus requires more restrictive rules, in terms of the relationship between the variables studied and the causal graph, in order to justify it as a tool for causal inference.

Secondly, once we have information transfer and flow in a common formalism, we find that

both appear to be susceptible to violations of the *stability* or *faithfulness* assumptions imported from causal Bayes net theory. As a result, both information transfer and flow are best seen as measures of complexity or epistemological clarity, though they are clearly distinct measures by virtue of the distinction between intervention and non-intervention on the causal graph. Similar effects are seen when we use transfer entropy as a time-series variant of information transfer.

The other proposal we have studied in the framework is convergent cross-mapping (CCM) (chapter 5). This requires more groundwork since its standard form is not presented in information-theoretic terms, instead using non-linear regression to evaluate a relationship. However, we show in chapter 5 that it can be stated in information-theoretic terms without losing very much. This allows us to give a direct comparison between CCM and transfer entropy.

CCM is based on a very different idea of how to measure causation than transfer entropy. Instead of using the causal Markov property – that effects are conditionally independent given their common causes – CCM uses the fact that when system A influences system B it is often possible to reconstruct the state of A from a temporal history of B. However, our analysis shows that in practice it is almost identical to a temporal mutual information, which is in turn very similar to transfer entropy without the requirement that common causes have been "conditioned out". The identification of this difference explains much of the practical difference and similarity between CCM and transfer entropy, and why they may appear to over- and under-estimate the presence of causal influences.

11.3 Hidden information as a feature of information dynamics

By *information dynamics* we refer to a view of systems as operating via transfer, storage and processing of information. Clearly here we have mainly been interested in the concept of information transfer, and from this perspective we suggest *hidden information transfer* as a notable feature of information dynamics. Loosely, we use this term to describe situations where information is transferred from some point A to another C, via some intermediary factor B, but where the transfer from A to B is less complex, or less easy to measure, than the transfer from A to C.

Though slightly counter-intuitive, this appears to be a common feature of practical communication systems, where successful transmission is dependent on the successful synchronisation of some ancillary components – the chaotic communication systems described in chapter 6, but also the models of embodied agents in chapters 8 and 10 seem to exhibit this type of information dynamics.

This is related to the concept of *synergy* in information transfer (Williams and Beer, 2010b), which refers to cases where conditioning on some variable actually increases the mutual informa-

tion:

$$I(A;C) < I(A;C|B)$$

This kind of relationship can seem counter-intuitive: if B and A are in some sense "informative" about C, and Alice already "knows" the state of B, but Bob does not, then it may seem that when they both learn about the state of A, Alice would gain less information about C than Bob does, because she already knows something about C, namely B. The problem with the intuition being that it is possible that A and B are only predictive of C *in combination* – you must know about both A *and* B, the knowledge of only A *or* B is useless (hence the term *synergy* – a classic example being the exclusive-or for Boolean variables: $C = (A \wedge \neg B) \vee (\neg A \wedge B)$).

Such synergetic relationships are likely to be the source of many examples of hidden information transfer, but synergy refers to a broader class, is does not necessarily relate to causal interpretations. We are particularly interested in the case of causal chains where the more remote parts of the chain have a more complex relationship than the closer parts.

This kind of dynamic would be problematic if information transfer is interpreted as strength of causation – since it seems implausible that a link could be stronger across the full chain that it is between intermediate links. But when adopting the inferential perspective that we have described, and in particular noting that low information transfer is not necessarily a reason to believe that a causal influence is absent, this result ceases to be problematic.

We have hypothesised that this phenomenon is likely to occur in cases where systems are strongly synchronised, since these are typical examples of strong causal influences that tend to lead to low information transfer. Specifically, we have looked at synchronisation in gait generation (chapter 10), but there may be other areas (see below, section 11.4.2) where we might expect to find similar phenomena for the same reasons.

11.3.1 Ergodicity in living and adaptive systems

This thesis has focused on stationary, and where possible (even if somewhat by contrivance) ergodic systems – that is, systems where we can (usually) analyse information flow by looking at time series data from a single run of a system. It has often been observed in the literature that such systems are not necessarily representative for a large class of phenomena that we might be interested in. The dynamics of the human brain (to take one example of many that we could find in living systems) in general most likely do not conform to the ergodic assumption, since it is subject to ongoing learning and adaptation, and generally does not behave repetitively over long periods. Information thoery originated in the study of communication systems where repetitive stochastic

(and hence ergodic) dynamics could be assumed, but even early proponents of the use of information theory in wider contexts (e.g. Wiener, 1965) were aware of the difficulty of applying it to non-ergodic systems.

It seems clear then, that for information theory to make a serious contribution in the study of living and adaptive systems in general that the problem of non-ergodicity must therefore be addressed. Since this thesis has targeted systems where we can find ergodicity it cannot claim to make a direct contribution in this direction, however there are two important points that can be derived from what has been addressed here:

- The consideration of exactly what contribution ergocity makes, and what problems arise when it fails, is naturally significant if one wishes to relax that assumption.
- The argument regarding the distinction between inference and complexity stands irrespective of whether we are studying ergodic or non-ergodic systems. The severity requirement can be applied provided a reasonable set of probabilistic entailments from causal hypotheses can be obtained.

11.4 Future directions

11.4.1 Inferring physical models

The emphasis in this thesis has been highly focussed on the relationship between *information* and causation. However, it seems clear that there are aspects of causation relevant to the design of autonomous robots that do not have to be construed in (exclusively) information theoretic terms – something which has received it seems relatively little attention here or in the wider literature.

In many discussions of causation in physical systems, the target is in fact the strength of *coupling* between continuously interacting components – the models used for testing e.g. by Sugihara et al. (2012) make it clear that physical coupling is what is meant by causation. If this is what we mean by a "causal" model, then inference of that model in effect means inference of the physical system itself.

Doing just that was the aim of the system proposed by Bongard and Lipson (2004) – here a physical model of a robot is "co-evolved" with the controller of the same robot – with the evolution of the controller targeted to improve the future evolution of the physical model. Much the same approach can be taken with more general physical models (not limited to robotics) (Bongard and Lipson, 2007).

What is striking about this approach, from our perspective, is that it bears a certain resemblance to the model of inference through *severe testing* discussed at various points through this thesis,

following the theory of Mayo (1996). Recall that in chapter 4, we argued that a causal influence can be inferred when it has passed a severe test: i.e. when we observe an information value that would be unlikely if the causal influence was absent.

The co-evolution approach of Bongard and Lipson (2004) consists of generating two "populations" – one population is a set of physical models of the robot (implemented in much the same way as our robot model in chapter 9, but across the population the shape of the robot is allowed to vary), the second population is a set of controllers, which dictate the robot's motor responses to sensor inputs.

First, one of the controllers is allowed to run on the real robot, the result being a data set (from sensor and motor recordings), which can be used to evolve the model population to better match the real robot. That is, the same controller is run on each member of the model population in a computer simulation, and data points from the simulation are compared to the real data collected from the robot. Those models which produced results more similar to the real robot are maintained in the population, whereas those which performed poorly are deleted and replaced with modified copies of the successful models. This step essentially keeps those models which accord well with the data – but it does not ensure that those models have been *severely* tested.

For that, a second stage of evolution is used, where the controller population is itself evolved against the current set of candidate models. That is, a controller is produced which maximises the differences in data produced by the different models in the population when they are simulated. After doing so, the first step (evolution of the models) will be repeated. This means that the controller is in effect optimized to produce a severe test of the models – the system aims to make the controller that will be run on the real robot one that would produce very different results across the different models specified by the model population. This means that when an improved model is selected by evolution, we can also say that model has been severely tested in the sense that if one of the other models in the population was closer to the true model, then we would have seen very different empirical results.

This thesis has argued that the actual value of information is not itself what is critical for making a causal inference: rather it is the severity with which a causal hypothesis is tested. However, we have not described ways in which severe tests could be achieved – the method described by Bongard and Lipson (2004) is an intriguing possible solution to this problem: simply evolve tests to maximise severity. This could be interpreted as evolving tests that maximise *disagreement* about mutual information values across candidate models, rather than focussing on the *absolute* value of mutual information.

11.4.2 Communication through coherence

We have often alluded to the capacity of synchronisation to act as a "carrier" or underlying mechanism to facilitate communication. In neuroscience, it has long been hypothesised that *neural synchrony* – the emergent synchronisation of firing patterns of neural groups – may play a role in modulating or facilitating information transfer among brain regions (Hoppensteadt and Izhikevich, 1998; Varela and Lachaux, 2001).

A particular current direction of research in neuroscience is the "communication through coherence" hypothesis (Fries, 2005). This suggests that information can be routed between regions of the brain by patterns of synchronisation, allowing for more precise selection of active information pathways than anatomical connectivity alone. Specifically, two synchronised brain regions periodically have a greater firing probability at the same time as each other. This means that information encoded in the spikes of one region have a better chance of influencing another region where the two are synchronised. When desynchronised, the incoherent firing patterns lead to non-overlapping periods of excitability in the neurons of each group, and thus information transfer is inhibited.

Akam and Kullmann (2014) cite this as one possible mechanism underlying *multiplexing* in neural systems – where multiple signals are transmitted through a single pathway and then later distinguished. This allows the brain to, for example, perform similar actions (such as reaching for an object) after selecting one of several possible information sources to base the action on (memory versus visual sensing of the object position for example).

The model of "hidden information transfer" relying on synchronised oscillators (chapter 6) would appear to have some relevance here. In neural systems, patterns of synchronisation are known to be task-dependent – Akam and Kullmann (2014) suggest that such changes in synchronisation modulate the *effective connectivity* of the system. Effective connectivity is often thought of as the pattern of "causal" influences in the neural system, as opposed to the physical structure (anatomical connectivity), or merely correlative associations (functional connectivity).

However, a precise characterisation of effective connectivity that accounts for features such as multiplexing may need further development. Certainly, transfer entropy and Granger causality have already been used as tools for measuring effective connectivity (Vicente et al., 2011), though there are also somewhat distinct mathematical definitions (Friston et al., 2003). The sense in which Akam and Kullmann (2014) use the term suggests high-level mapping, such as from sensory inputs to motor or cognitive aspects over which they exert some control – the patterns of synchronisation determine the causal pathways, but are not themselves instances of the kind of causal influence intended, they are auxiliary to the primary information transfer in the same sense that the chaotic

oscillators of Cuomo and Oppenheim (1993) discussed in chapter 6 facilitate information transfer. Indeed, as we have discussed, strong synchrony is the kind of situation where we expect low information transfer – but such strong synchrony may be indicative of the *routing* of effective connectivity through the system as a whole. A robotic system implementing such task-dependent routing of causal influences (possibly utilising neural models for control) would be an illuminating way to approach this question.

11.5 Closing remarks

The epigraph from chapter 1 quoted Claude Shannon, one of the originators of modern information theory, from an editorial he wrote entitled "The Bandwagon" (Shannon, 1956). In it, Shannon seems concerned about contemporary translations of information-theoretic tools into sundry scientific fields without, in his view, sufficient rigour.

The field has moved on since the 1950s and much of the ground work that might have been worryingly lacking then has been substantially improved since. However, Shannon's reservations still suggest some caution is apt, as he notes: "if, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact."

Thus, central to our argument has been the way in which information-theoretic statistics are used to make inferences, from experiment, about causation. The concept of *severe testing* which has been much discussed addresses the need for some reasoning to underpin the way we make experimental inferences in spite of the problem of induction – the inability to go, logically, from a collection of singular experimental results to a universal generalisation. Recall this is much the same problem that led Hume to conclude that causation cannot be directly observed. This has been central to our recapitulation of information transfer as a descriptor of epistemological clarity.

From this starting point we have developed the idea of hidden information, and analysed the information dynamics of some robotic systems. It seems clear from our results that although there is a connection between the causal influences and information transfer, it is not always a trivial one. There is something more. As suggested by Lizier and Prokopenko (2010), perhaps it would be best to think of information transfer as describing computation, but precisely what we mean by this seems to be something that still requires refinement. This thesis has aimed to make some progress in this direction.

Bibliography

Akam, T. and Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience*, 15(2):111–22. Cited on 234

Akobeng, A. K. (2005). Principles of evidence based medicine. *Archives of disease in childhood*, 90(8):837–40. Cited on 16

Amblard, P.-O. and Michel, O. J. J. (2011). On directed information theory and Granger causality graphs. *Journal of Computational Neuroscience*, 30(1):7–16. Cited on 95

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683. Cited on 50, 51

Arkin, R. C. (1998). *Behaviour-based robotics*. MIT Press, Cambridge, Mass. Cited on 3

Arntzenius, F. (1992). The common cause principle. In *PSA 1992: Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association*, volume 2, pages 227–237. Cited on 31, 33

Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B*, 63(3):329–339. Cited on 121

Ay, N. and Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, 11(01):17. Cited on 7, 9, 12, 67, 69, 70, 82, 87, 89, 95, 97, 98, 100, 119, 149, 229

Baptista, M. and Kurths, J. (2005). Chaotic channel. *Physical Review E*, 72(4):1–4. Cited on 96

Barnett, L. (2009). Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters*, 103(23):1–10. Cited on 88, 102, 111, 149

Beer, R., Chiel, H., Quinn, R., and Larsson, P. (1992). A distributed neural network architecture for hexapod robot locomotion. *Neural Computation*, 4(3):356–365. Cited on 144, 161

Beer, R. D. (1995). A dynamical systems perspective interaction on agent-environment interaction. *Artificial Intelligence*, 72(1–2):173–215. Cited on 4, 161

Belyakov, A. O., Seyranian, A. P., and Luongo, A. (2009). Dynamics of the pendulum with periodically varying length. *Physica D: Nonlinear Phenomena*, 238(16):1589–1597. Cited on 150

Bennett, M., Schatz, M. F., Rockwood, H., and Wiesenfeld, K. (2002). Huygens's clocks. *Proceedings of the Royal Society A*, 458(2019):563–579. Cited on 70

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: an information theoretic perspective. *Bio Systems*, 91(2):331–45. Cited on 119, 148

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–63. Cited on 1, 96, 121

Blickhan, R., Seyfarth, A., Geyer, H., Grimmer, S., Wagner, H., and Günther, M. (2007). Intelligence by mechanics. *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences*, 365(1850):199–220. Cited on 141

Bongard, J. and Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24):9943–9948. Cited on 232

Bongard, J. C. and Lipson, H. (2004). Once More Unto the Breach: Co-evolving a robot and its simulator. In *ALIFE IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. Cited on 232, 233

Boyle, J. H. (2009). *C. elegans locomotion: an integrated approach*. PhD thesis, University of Leeds. Cited on 144

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, Mass. Cited on 2, 22

Breiman, L. (1969). *Proability and Stochastic Processes: With a View Toward Applications*. Houghton Mifflin Company, Boston, Mass. Cited on 89, 95, 154

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23. Cited on 2, 3, 151

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1-2):3–15. Cited on 4

Buchli, J., Iida, F., and Ijspeert, A. (2006). Finding resonance: adaptive frequency oscillators for dynamic legged locomotion. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3903–3909. IEEE. Cited on 144

Campos, R., Matos, V., and Santos, C. (2010). Hexapod locomotion: A nonlinear dynamical systems approach. *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, pages 1546–1551. Cited on 144

Cartwright, N. (1994). *Nature's capacities and their measurement*. Oxford University Press, Oxford. Cited on 7, 28

Cartwright, N. (2007a). Are RCTs the gold standard? *BioSocieties*, 2(1):11–20. Cited on 17, 28

Cartwright, N. (2007b). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, Cambridge. Cited on 47, 50

Cartwright, N. and Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, 16(2):260–6. Cited on 17, 24

Ceguerra, R. V., Lizier, J. T., and Zomaya, A. Y. (2011). Information storage and transfer in the synchronization process in locally-connected networks. In *Proceedings of the 2011 IEEE Symposium on Artificial Life*. Cited on 54, 59, 61, 96

Chemero, A. (2009). *Radical Embodied Cognitive Science*. MIT Press, Cambridge, MA. Cited on 4, 6

Clark, A. and Toribio, J. (1994). Doing without representing. *Synthese*, 101. Cited on 4

Cohen, I. (2006). Immune system computation and the immunological homunculus. *Model Driven Engineering Languages and Systems*, pages 499–512. Cited on 96

Collins, J. J. and Stewart, I. (1993). Coupled nonlinear oscillators and the symmetries of animal gaits. *Journal of Nonlinear Science*, 3(1):349–392. Cited on 4, 54, 144, 147

Collins, S., Ruina, A., Tedrake, R., and Wisse, M. (2005). Efficient bipedal robots based on passive-dynamic walkers. *Science*, 307(5712):1082–5. Cited on 142

Collins, S. H., Wisse, M., and Ruina, A. (2001). A three-dimensional passive-dynamic walking robot with two legs and knees. *The International Journal of Robotics Research*, 20(7):607–615. Cited on 137

Colquhoun, D. (2011). In praise of randomisation. In Dawid, P., Twining, W., and Vasilaki, M., editors, *Evidence, Inference and Enquiry*, page 323. Oxford University Press, Oxford. Cited on 16

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97. Cited on 6

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Jon Wiley & Sons, Hoboken, NJ, 2nd edition. Cited on 1, 107, 108, 109, 110, 120, 123

Craig, J. J. (2005). *Introduction to robotics: mechanics and control*. Pearson/Prentice Hall, Upper Saddle River, NJ. Cited on 2

Cruse, H., Arena, P., and Frasca, M. (2002). Cellular nonlinear network-based bio-inspired decentralized control of locomotion for hexapod robots. *Adaptive Behavior*, 10(2):97–111. Cited on 10, 144, 161, 172

Cruse, H., Dürr, V., and Schmitz, J. (2007). Insect walking is based on a decentralized architecture revealing a simple and robust controller. *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences*, 365(1850):221–50. Cited on 144, 161, 172, 173

Crutchfield, J. and Young, K. (1989). Inferring statistical complexity. *Physical Review Letters*, 63(2):105–108. Cited on 1, 96

Crutchfield, J. P. and Mitchell, M. (1995). The evolution of emergent computation. *Proceedings of the National Academy of Sciences of the United States of America*, 92(23):10742–6. Cited on 96

Cuomo, K. and Oppenheim, A. (1993). Circuit implementation of synchronized chaos with applications to communications. *Physical Review Letters*, 71(1):65–68. Cited on 128, 160, 235

Dabrowska, D. and Speed, T. (1990). On the application of probability theory to agricultural experiments: Essay on principles. *Statistical Science*, 5(4):465–472. Cited on 19

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424. Cited on 18, 21

Dawid, A. P. (2009). Beware of the DAG! In *JMLR: Workshop and Conference Proceedings 6*, pages 59–86. Cited on 7, 47

Der, R., Güttler, F., and Ay, N. (2008). Predictive information and emergent cooperativity in a chain of mobile robots. In *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*. MIT Press. Cited on 1, 6, 121

Der, R., Hesse, F., and Martius, G. (2006). Rocking Stamper and Jumping Snakes from a Dynamical Systems Approach to Artificial Life. *Adaptive Behavior*, 14(2):105–115. Cited on 121

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, 11(3):189–228. Cited on 199

Edelman, G. M. and Gally, J. a. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13763–8. Cited on 26

Eichler, M. (2012a). Causal inference in time series analysis. In Berzuini, C., Dawid, P., and Bernadinelli, L., editors, *Causality: Statistical perspectives and applications*, chapter 22, pages 327–354. Wiley, Hoboken, NJ. Cited on 95

Eichler, M. (2012b). Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2):233–268. Cited on 95

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, 4(5):41–60. Cited on 96

Evans, D. (2008). A computationally efficient estimator for mutual information. *Proceedings of the Royal Society A*, 464(2093):1203–1215. Cited on 155

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1–40. Cited on 35

Fisher, R. (1935). *The Design of Experiments*. Oliver & Boyd, Oxford. Cited on 16

Fisher, R. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):69–78. Cited on 74

Floreano, D., Husbands, P., and Nolfi, S. (2008). Evolutionary robotics. In Siciliano, B. and Khatib, O., editors, *Springer Handbook of Robotics*, pages 1423–1451. Springer Berlin Heidelberg. Cited on 3, 136

Foster, D. V. and Grassberger, P. (2011). Lower bounds on mutual information. *Physical Review E*, 83:010101(R). Cited on 107

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, 9(10):474–80. Cited on 234

Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302. Cited on 234

Fukuoka, Y., Kimura, H., and Cohen, A. (2003). Adaptive dynamic walking of a quadruped robot on irregular terrain based on biological concepts. *The International Journal of Robotics Research*, 22(3-4):187. Cited on 144

Gabrielli, G. and von Kármán, T. (1950). What price speed? *Mechanical Engineering*, 72:775–781. Cited on 142

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, 20(5):507–534. Cited on 41

Geyer, H., Seyfarth, A., and Blickhan, R. (2006). Compliant leg behaviour explains basic dynamics of walking and running. *Proceedings of the Royal Society B*, 273(1603):2861–7. Cited on 141

Golubitsky, M. (1998). A modular network for legged locomotion. *Physica D: Nonlinear Phenomena*, 115(1-2):56–72. Cited on 144

Golubitsky, M., Stewart, I., Buono, P. L., and Collins, J. J. (1999). Symmetry in locomotor central pattern generators and animal gaits. *Nature*, 401(6754):693–5. Cited on 144

Good, I. J. (1961a). A causal calculus (I). *The British journal for the philosophy of science*, 11(44):305–318. Cited on 27

Good, I. J. (1961b). A causal calculus (II). *The British journal for the philosophy of science*, 12(45):43–51. Cited on 27

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438. Cited on 1, 27, 102, 104, 110, 117

Gregorio, P., Ahmadi, M., and Buehler, M. (1997). Design, control, and energetics of an electrically actuated legged robot. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, 27(4):626–34. Cited on 142, 143

Hacking, I. (1988). Telepathy: origins of randomization in experimental design. *Isis*, 79(3):427–451. Cited on 16

Harvey, I., Di Paolo, E., Wood, R., Quinn, M., and Tuci, E. (2005). Evolutionary robotics: a new scientific tool for studying cognition. *Artificial life*, 11(1-2):79–98. Cited on 3, 4

Harvey, I., Vaughan, E., and Di Paolo, E. (2004). Time and Motion Studies: The Dynamics of Cognition, Computation and Humanoid Walking. In *HART 2004. Fourth Internaional Symposium on Human and Artificial Intelligence Systems: From Control to Autonomy*. Cited on 137, 138, 139, 141, 144

Haynes, L., Service, O., Goldacre, B., and Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Technical report, UK Cabinet Office. Cited on 16

Henderson, M. (2012). *The Geek Manifesto*. Transworld Publishers, London. Cited on 16

Hesslow, G. (1976). Discussion: two notes on the probabilistic approach to causality. *Philosophy of Science*, 43(2):290–292. Cited on 44

Hindmarsh, A. (1983). ODEPACK, a systematized collection of ODE solvers. In Stepleman, R. S., editor, *Scientific Computing*, pages 55–64. North-Holland, Amsterdam. Cited on 58

Hirose, M. and Ogawa, K. (2007). Honda humanoid robots development. *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences*, 365(1850):11–9. Cited on 137, 139

Hitchcock, C. (2012). Probabilistic Causation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition. Cited on 31

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960. Cited on 7, 19, 27, 104, 111

Hoover, K. (2003). Nonstationary time series, cointegration, and the principle of the common cause. *The British Journal for the Philosophy of Science*, 54(4):527–551. Cited on 34, 103

Hoppensteadt, F. C. and Izhikevich, E. M. (1998). Thalamo-cortical interactions modeled by weakly connected oscillators: could the brain use FM radio principles? *Bio Systems*, 48(1-3):85–94. Cited on 234

Hume, D. (2006). *An Enquiry Concerning Human Understanding*. Project Gutenberg. Cited on 18

I-Bioloid (2010). Bioloid robot brackets. `http://www.thingiverse.com/thing:5192` [Accessed: 10th May 2014]. Cited on 164

Iida, F., Gomez, G., and Pfeifer, R. (2005). Exploiting body dynamics for controlling a running quadruped robot. In *ICAR '05. Proceedings of the 12th International Conference on Advanced Robotics*, pages 229–235. IEEE. Cited on 141

Ijspeert, A. J. (2008). Central pattern generators for locomotion control in animals and robots: a review. *Neural Networks*, 21(4):642–53. Cited on 4, 143

Ijspeert, A. J., Crespi, A., Ryczko, D., and Cabelguen, J.-M. (2007). From swimming to walking with a salamander robot driven by a spinal cord model. *Science*, 315(5817):1416–20. Cited on 144, 146

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schlkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358. Cited on 7, 149

Jung, T., Polani, D., and Stone, P. (2011). Empowerment for continuous agent–environment systems. *Adaptive Behavior*, 19(1):16–39. Cited on 138

Kajita, S. and Espiau, B. (2008). Legged Robots. In Siciliano, B. and Khatib, O., editors, *Springer Handbook of Robotics*, pages 361–389. Springer Berlin Heidelberg. Cited on 136, 139, 140, 141

Kantz, H. and Schreiber, T. (2003). *Nonlinear Time Series Analysis*. Cambridge University Press, 2nd edition. Cited on 90, 91, 95

Kiers, K., Schmidt, D., and Sprott, J. C. (2004). Precision measurements of a simple chaotic circuit. *American Journal of Physics*, 72(4):503. Cited on 56, 57

Kim, M., Kim, I., Park, S., and Oh, J. (2008). Realization of stretch-legged walking of the humanoid robot. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, pages 118–124. IEEE. Cited on 137

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9):3354–9. Cited on 107

Kitano, H. (2002). Computational systems biology. *Nature*, 420(6912):206–210. Cited on 96

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837. Cited on 26

Klyubin, A., Polani, D., and Nehaniv, C. (2005). Empowerment: A Universal Agent-Centric Measure of Control. In *2005 IEEE Congress on Evolutionary Computation*, pages 128–135. IEEE. Cited on 1, 6, 138

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2008). Keep your options open: an information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12):e4018. Cited on 96, 148

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):16. Cited on 114, 155, 197, 201

Kuhlmann, M. (2011). Mechanisms in Dynamically Complex Systems. In McKay Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*. Oxford University Press. Cited on 17

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley and Sons. Cited on 12, 75

Kuniyoshi, Y. and Suzuki, S. (2004). Dynamic emergence and adaptation of behavior through embodiment as coupled chaotic field. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. Cited on 145

Ladyman, J., Lambert, J., and Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67. Cited on 17

LaValle, S. M. (2011). Motion Planning. *IEEE Robotics & Automation Magazine*, 18(1):79–89. Cited on 2

Lewis, D. (1986). Events. In *Philosophical Papers II*. Oxford University Press, New York. Cited on 33

Lewis, M. A. and Bekey, G. A. (2002). Gait adaptation in a quadruped robot. *Autonomous Robots*, 12(3):301–312. Cited on 144

Lewis, M. A., Etienne-Cummings, R., Hartmann, M. J., Xu, Z. R., and Cohen, A. H. (2003). An in silico central pattern generator: silicon oscillator, coupling, entrainment, and physical computation. *Biological cybernetics*, 88(2):137–51. Cited on 138, 146

Lizier, J. T. (2010). *The local information dynamics of distributed computation in complex systems*. PhD thesis, The University of Sydney. Cited on 2, 7, 96

Lizier, J. T. and Prokopenko, M. (2010). Differentiating information transfer and causal effect. *The European Physical Journal B*, 73(4):605–615. Cited on 55, 67, 86, 96, 117, 149, 235

Lungarella, M. and Sporns, O. (2006). Mapping information flow in sensorimotor networks. *PLoS Computational Biology*, 2(10):e144. Cited on 7, 55, 63, 96, 119, 148

Marschinski, R. and Kantz, H. (2002). Analysing the information flow between financial time series. *The European Physical Journal B*, 30(2):275–281. Cited on 7, 55, 60, 63, 197, 201

Matsuoka, K. (1985). Sustained oscillations generated by mutually inhibiting neurons with adaptation. *Biological cybernetics*, 52(6):367–76. Cited on 143

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press. Cited on 71, 125, 196, 228, 233

Mayo, D. G. and Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. *IMS Lecture Notes - Monographs Series*, 49:77–97. Cited on 196

Mayo, D. G. and Spanos, A. (2010). Error statistics. In *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 1–46. Cited on 72, 73, 74, 77

McGeer, T. (1990). Passive Dynamic Walking. *The International Journal of Robotics Research*, 9(2):62–82. Cited on 60, 136, 139

Mead, C. (1990). Neuromorphic Electronic Systems. *Proceedings of the IEEE*, 78(10):1629–1636. Cited on 146

Menzies, P. (2009). Counterfactual Theories of Causation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2009 edition. Cited on 18

Mill, J. S. (1843). *A System of Logic*. Longmans, Green and Co., London. Cited on 18, 24

Mitchell, M. (2006). Complex systems: Network thinking. *Artificial Intelligence*, 170(18). Cited on 96

Mitchell, S. D. (2009). *Unsimple Truths: Science, Complexity, and Policy*. University of Chicago Press, Chicago. Cited on 26

Moioli, R. C., Vargas, P. A., and Husbands, P. (2012). Synchronisation effects on the behavioural performance and information dynamics of a simulated minimally cognitive robotic agent. *Biological cybernetics*, 106(6-7):407–427. Cited on 7, 119, 148

Morris, B., Westervelt, E., Chevallereau, C., Buche, G., and Grizzle, J. (2006). Achieving bipedal running with RABBIT: Six steps toward infinity. In *Fast Motions in Biomechanics and Robotics*, volume 340, pages 277–297. Springer. Cited on 141

Morris, W. E. (2013). David Hume. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 201 edition. Cited on 18

Murphy, K. (1998). A Brief Introduction to Graphical Models and Bayesian Networks. `http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html`. Accessed: 10th December 2013. Cited on 35

Murphy, K. N. and Raibert, M. H. (1985). Trotting and Bounding in a Planar Two-Legged Model. In Morecki, A., Bianchi, G., and Kędzior, K., editors, *Theory and Practice of Robots and Manipulators. Proc. of RoManSy '84: The Fifth CISM-IFToMM Symposium*, pages 411–420, London. Kogan Page. Cited on 141

Neyman, J. (1923). On The Application of Probability Theory to Agricultural Experiments. Essay on Principles. *Roczniki Nauk Rolniczych*, X:1–51. Cited on 19

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society. Series A, Containing Papers of a Mathematical or Physical Character*, 231(1933):289–337. Cited on 76

Nurse, P. (2008). Life, logic and information. *Nature*, 454(7203):424–6. Cited on 96

Ogura, Y., Shimomura, K., Kondo, H., Morishima, A., Okubo, T., Momoki, S., Lim, H., and Takanishi, A. (2006). Human-like Walking with Knee Stretched, Heel-contact and Toe-off Motion by a Humanoid Robot. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3976–3981. IEEE. Cited on 137

Oka, M. and Ikegami, T. (2012). Characterizing Autonomy in the Web via Transfer Entropy Network. *Artificial Life 13*, pages 234–242. Cited on 96

Pearl, J. (2000). The logic of counterfactuals in causal inference (Discussion of 'Causal inference without counterfactuals' by A.P. Dawid). *Journal of the American Statistical Association*, 95(450):428–435. Cited on 21

Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition. Cited on 7, 23, 28, 35, 36, 39, 40, 44, 45, 48, 49, 52, 78, 79, 80, 81, 95, 99, 104, 229

Pearson, K. (1911). *Grammar of Science*. A. and C. Black Publishers, London, 3rd edition. Cited on 21

Pearson, K. (1995). Proprioceptive regulation of locomotion. *Current Opinion in Neurobiology*, 5(6):786–791. Cited on 144

Pecora, L. and Carroll, T. (1990). Synchronization in chaotic systems. *Physical Review Letters*, 64(8):821–824. Cited on 128, 145

Pfeifer, R. and Bongard, J. (2007). *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, Cambirdge, Mass. Cited on 119, 142

Pfeifer, R. and Iida, F. (2004). Embodied artificial intelligence: Trends and challenges. In Iida, F., Pfeifer, R., Steels, L., and Kuniyoshi, Y., editors, *Embodied Artificial Intelligence*, volume 3139 of *Lecture Notes in Computer Science*, pages 1–26. Springer Berlin Heidelberg. Cited on 161

Pfeifer, R., Iida, F., and Gómez, G. (2006). Morphological computation for adaptive behavior and cognition. *International Congress Series*, 1291:22–29. Cited on 6

Pfeifer, R., Lungarella, M., and Iida, F. (2007a). Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088–93. Cited on 152

Pfeifer, R., Lungarella, M., Sporns, O., and Kuniyoshi, Y. (2007b). On the information theoretic implications of embodiment - principles and methods. In Lungarella, M., Iida, F., Bongard, J., and Pfeifer, R., editors, *50 Years of Artificial Intelligence*, volume 4850 of *Lecture Notes in Computer Science*, pages 76–86, Berlin / Heidelberg. Springer. Cited on 96, 148

Pikovsky, A., Rosenblum, M., and Kurths, J. (2001). *Synchronization: a universal concept in nonlinear sciences*. Cambridge University Press, Cambridge. Cited on 4, 54, 59, 60, 71, 92

Pinsky, M. and Zevin, A. (1999). Oscillations of a pendulum with a periodically varying length and a model of swing. *International Journal of Non-Linear Mechanics*, 34(1):105–109. Cited on 150

Pitti, A., Lungarella, M., and Kuniyoshi, Y. (2009). Generating spatiotemporal joint torque patterns from dynamical synchronization of distributed pattern generators. *Frontiers in Neurorobotics*, 3(2). Cited on 55, 56, 59, 60, 63, 96, 144, 145, 148

Pitti, A., Niiyama, R., and Kuniyoshi, Y. (2010). Creating and modulating rhythms by controlling the physics of the body. *Autonomous Robots*, 28(3):317–329. Cited on 145

Politis, D. and Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313. Cited on 200

Raibert, M., Blankespoor, K., Nelson, G., Playter, R., and Others (2008). Bigdog, the rough-terrain quadruped robot. In *Proceedings of the 17th International Federation of Automation Control*, pages 10822–10825. Cited on 141

Raibert, M. H. (1986). *Legged Robots that Balance*. MIT Press, Cambridge, MA. Cited on 140, 141, 144

Raibert, M. H. and Brown, H. B. (1984). Experiments in balance with a 2D one-legged hopping machine. *ASME Journal of Dynamic Systems, Measurement, and Control*, 106:75–81. Cited on 140

Raibert, M. H., Brown, H. B., and Chepponis, M. (1984). Experiments in Balance with a 3D One-Legged Hopping Machine. *The International Journal of Robotics Research*, 3(2):75–92. Cited on 140

Raibert, M. H., Chepponis, M., and Brown, H. B. (1986). Running on four legs as though they were one. *IEEE Journal of Robotics and Automation*, 2(2):70–82. Cited on 140

Reichenbach, H. (1956). *The Direction of Time*. University of California Press, Berkeley, Los Angeles. Cited on 7, 27, 28

Remy, C. D., Buffinton, K., and Siegwart, R. (2009). Stability Analysis of Passive Dynamic Walking of Quadrupeds. *The International Journal of Robotics Research*, 29(9):1173–1185. Cited on 137, 143

Ringrose, R. (1997). *Self-stabilizing running*. PhD thesis, Massachusetts Institute of Technology. Cited on 141

Rissanen, J. (1986). Stochastic Complexity and Modeling. *The Annals of Statistics*, 14(3):1080–1100. Cited on 96

Rissanen, J. (1987). Measures of mutual and causal dependence between two time series (Corresp.). *IEEE Transactions on Information Theory*, 33(4):598–601. Cited on 95

Robotis (2006). *Dynamixel AX-12 User's Manual*. Cited on 165, 166, 167

Rosenblum, M., Pikovsky, A., and Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Physical Review Letters*, 76(11):1804–1807. Cited on 54, 145

Rosvall, M. and Bergstrom, C. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7327–7331. Cited on 96

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(688-701). Cited on 19

Runge, J., Heitzig, J., Marwan, N., and Kurths, J. (2012a). Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E*, 86:061121. Cited on 119

Runge, J., Heitzig, J., Petoukhov, V., and Kurths, J. (2012b). Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical Review Letters*, 108:258701. Cited on 7, 95, 96, 117, 119

Sabourin, C. and Bruneau, O. (2005). Robustness of the dynamic walk of a biped robot subjected to disturbing external forces by using CMAC neural networks. *Robotics and Autonomous Systems*, 51(2-3):81–99. Cited on 141

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72. Cited on 16

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton. Cited on 95

Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press, New York. Cited on 95

Sauer, T. (2006). Attractor Reconstruction. *Scholarpedia*, 1(10):1727. Cited on 90

Sauer, T. (2012). Numerical solution of stochastic differential equations in finance. In Duan, J.-C., Härdle, W. K., and Gentle, J. E., editors, *Handbook of Computational Finance*, pages 529–550. Springer, Berlin. Cited on 92, 154

Sauer, T. and Yorke, J. (1991). Embedology. *Journal of Statistical Physics*, 65(3):579–616. Cited on 91

Scarfogliero, U., Stefanini, C., and Dario, P. (2009). The use of compliant joints and elastic energy storage in bio-inspired legged robots. *Mechanism and Machine Theory*, 44(3):580–590. Cited on 141

Schäfer, C., Rosenblum, M. G., Kurths, J., and Abel, H. H. (1998). Heartbeat synchronized with ventilation. *Nature*, 392(6673):239–240. Cited on 55, 61, 62

Schilling, M., Hoinville, T., Schmitz, J., and Cruse, H. (2013). Walknet, a bio-inspired controller for hexapod walking. *Biological Cybernetics*, 107(4):397–419. Cited on 144, 161

Schmidt, N. M., Hoffmann, M., Nakajima, K., and Pfeifer, R. (2012). Bootstrapping Perception Using Information Theory: Case Studies in a Quadruped Robot Running on Different Grounds. *Advances in Complex Systems*, 16:1250078. Cited on 7, 148, 221

Schreiber, T. (1990). Spatio-temporal structure in coupled map lattices: two-point correlations versus mutual information. *Journal of Physics A*, 23(8):L393–L398. Cited on 110

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2):461–464. Cited on 1, 55, 88, 102, 110, 119, 149, 229

Seth, A. K. (2007). Granger causality. *Scholarpedia*, 2(7):1667. Cited on 27

Seth, A. K., Barrett, A. B., and Barnett, L. (2011). Causal density and integrated information as measures of conscious level. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 369(1952):3748–67. Cited on 7

Shalizi, C. R. (2003). Optimal nonlinear prediction of random fields on networks. In *Discrete Models for Complex Systems, DMCS'03, Lyon, France, June 16-19, 2003*, pages 11–30. Cited on 86

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:329–423. Cited on 2, 12, 13, 95

Shannon, C. E. (1956). The Bandwagon. *IRE Transactions on Information Theory*, 2(1):3. Cited on 1, 235

Shim, Y. and Husbands, P. (2010). Chaotic Search of Emergent Locomotion Patterns for a Bodily Coupled Robotic System. In *Artificial Life XII: Proceedings of the Twelfth International Conference on the Simulation and Synthesis of Living Systems*, pages 757–764. Cited on 145

Simon, H. (1977). Causal ordering and identifiability. In *Models of Discovery*, volume 54 of *Boston Studies in the Philosophy of Science*, pages 53–80. Springer Netherlands. Cited on 18

Smirnov, D. A. (2013). Spurious causalities with transfer entropy. *Physical Review E*, 87(4):042917. Cited on 117

Sober, E. (1984). Common cause explanation. *Philosophy of Science*, 51:212–241. Cited on 7, 33, 89

Sober, E. (2001). Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *The British Journal for the Philosophy of Science*, 52(2):331–346. Cited on 7, 33, 89

Sokal, R. R. and Rohlf, F. J. (1991). *Biometry*. W. H. Freeman, 3rd edition. Cited on 74

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition. Cited on 7, 35, 44, 45, 48, 49, 78, 95

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In Galavotti, M. C., Suppes, P., and Constantini, D., editors, *Stochastic Dependence and Causality*, chapter 9, pages 157–172. University of Chicago Press, Chicago. Cited on 51

Sprague, J. M. (1966). Interaction of cortex and superior colliculus in mediation of visually guided behavior in the cat. *Science*, 153(3743):1544–7. Cited on 26

Sprott, J. C. (2000). Simple chaotic systems and circuits. *American Journal of Physics*, 68(8):758. Cited on 56

Staniek, M. and Lehnertz, K. (2008). Symbolic Transfer Entropy. *Physical Review Letters*, 100(15):1–4. Cited on 111, 155

Stepney, S., Braunstein, S., Clark, J., Tyrrell, A., Adamatzky, A., Smith, R., Addis, T., Johnson, C., Timmis, J., Welch, P., Milner, R., and Partridge, D. (2005). Journeys in non-classical computation I: A grand challenge for computing research. *International Journal of Parallel, Emergent and Distributed Systems*, 20(1):5–19. Cited on 138

Stigler, S. (1978). Mathematical statistics in the early states. *The Annals of Statistics*, 8(2):239–265. Cited on 16

Still, S., Hepp, K., and Douglas, R. J. (2006). Neuromorphic walking gait control. *IEEE Transactions on Neural Networks*, 17(2):496–508. Cited on 147

Strogatz, S. H. (2000). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview Press, Cambridge, Mass. Cited on 13

Strogatz, S. H. (2004). *Sync: The emerging science of spotaneous order*. Penguin Books, London. Cited on 4, 71

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338(6106):496–500. Cited on 7, 9, 101, 102, 104, 105, 111, 114, 116, 117, 119, 229, 232

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North Holland Publishing Co. Cited on 27

Taga, G. and Yamaguchi, Y. (1991). Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment. *Biological Cybernetics*, 65(3):147–159. Cited on 143, 145, 146

Takanishi, A., Ishida, M., Yamazaki, Y., and Kato, I. (1985). The realization of dynamic walking by the biped walking robot WL-10RD. In *International Conference on Advanced Robotics (ICAR85)*, pages 459–466. Cited on 139

Takens, F. (1981). Detecting strange attractors in turbulence. In Rand, D. and Young, L.-S., editors, *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer Berlin / Heidelberg. Cited on 104

Thorniley, J. (2011). An improved transfer entropy method for establishing causal effects in synchronizing oscillators. In Lenaerts, T., Giacobini, M., Bersini, H., Bourgine, P., Dorigo, M., and Doursat, R., editors, *ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*. MIT Press. Cited on 54

Thorniley, J. and Husbands, P. (2013). Hidden information transfer in an autonomous swinging robot. In *Advances in Artificial Life, ECAL 2013*, volume 12, pages 513–520. MIT Press. Cited on 148

Touchette, H. and Lloyd, S. (1999). Information-Theoretic Limits of Control. *Physical Review Letters*, 84(6):4. Cited on 6

Tucker, V. A. (1970). Energetic cost of locomotion in animals. *Comparative biochemistry and physiology*, 34(4):841–6. Cited on 142

Twyman, R. (2004). A brief history of clinical trials. `http://genome.wellcome.ac.uk/doc_WTD020948.html`. Accessed: 10th December 2013. Cited on 16

Van Gelder, T. (1995). What might cognition be, if not computation. *The Journal of Philosophy*, 92(7):345–381. Cited on 4

Varela, F. and Lachaux, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2:229–239. Cited on 234

Vaughan, E., Di Paolo, E., and Harvey, I. (2004). The evolution of control and adaptation in a 3D powered passive dynamic walker. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R., editors, *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 139–145. MIT Press. Cited on 137

Vaughan, E., Di Paolo, E., and Harvey, I. (2005). The tango of a load balancing biped. In Armada, M. and Gonzalez De Santos, P., editors, *Climbing and Walking Robots, Proceedings of the 7th International Conference CLAWAR 2004*, Berlin. Springer. Cited on 137

Vaughan, E., Di Paolo, E., and Harvey, I. (2014). Incremental Evolution of an Omni-directional Biped for Rugged Terrain. In Vargas, P., Di Paolo, E., Harvey, I., and Husbands, P., editors, *The Horizons of Evolutionary Robotics*, pages 237–278, Cambridge, Mass. MIT Press. Cited on 137

Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67. Cited on 7, 117, 197, 234

Vlachos, I. and Kugiumtzis, D. (2010). Nonuniform state-space reconstruction and coupling detection. *Physical Review E*, 82(1):016207. Cited on 114, 197

Vogelstein, R. J., Tenore, F. V. G., Guevremont, L., Etienne-Cummings, R., and Mushahwar, V. K. (2008). A Silicon Central Pattern Generator Controls Locomotion in Vivo. *IEEE Transactions on Biomedical Circuits and Systems*, 2(3):212–222. Cited on 147

Vukobratović, M. and Borovac, B. (2004). Zero-moment point – thirty five years of its life. *International Journal of Humanoid Robotics*, 1(1):157–173. Cited on 139

Wiener, N. (1965). *Cybernetics*. MIT Press, Cambridge, MA, 2nd edition. Cited on 95, 232

Wiggins, S. (2003). *Introduction to applied nonlinear dynamical systems and chaos*. Springer-Verlag, New York, 2nd edition. Cited on 138

Williams, P. and Beer, R. D. (2010a). Information Dynamics of Evolved Agents. In *From Animals to Animats 11*, pages 38–49. Springer, Berlin / Heidelberg. Cited on 61, 96, 119, 148

Williams, P. L. and Beer, R. D. (2010b). Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515. Cited on 159, 230

Winship, C. and Sobel, M. (2004). Causal Inference in Sociological Studies. In Hardy, M. and Bryman, A., editors, *Handbook of Data Analysis*. SAGE Publications. Cited on 18, 19

Woodward, J. (2004). *Making Things Happen*. Oxford University Press. Cited on 7, 16, 18

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585. Cited on 34

Zahedi, K. and Ay, N. (2013). Quantifying Morphological Computation. *Entropy*, 15(5):1887–1915. Cited on 7

Zahedi, K., Ay, N., and Der, R. (2010). Higher Coordination With Less Control–A Result of Information Maximization in the Sensorimotor Loop. *Adaptive Behavior*, 18(3-4):338–355. Cited on 6

Zevin, A. and Filonenko, L. (2007). A qualitative investigation of the oscillations of a pendulum with a periodically varying length and a mathematical model of a swing. *Journal of Applied Mathematics and Mechanics*, 71(6):892–904. Cited on 150

Zhang, J. and Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference. *Minds and Machines*, 18(2):239–271. Cited on 48