Terry Bossomaier · Lionel Barnett
Michael Harré · Joseph T. Lizier

An Introduction to Transfer Entropy

Information Flow in Complex Systems

Springer

An Introduction to Transfer Entropy

Terry Bossomaier · Lionel Barnett
Michael Harré · Joseph T. Lizier

An Introduction to Transfer Entropy

Information Flow in Complex Systems

Springer

Terry Bossomaier
School of Computing and Mathematics
Charles Sturt University
Bathurst, NSW
Australia

Lionel Barnett
Department of Informatics
University of Sussex
Brighton
UK

Michael Harré
Department of Civil Engineering
University of Sydney
Darlington, NSW
Australia

Joseph T. Lizier
Department of Civil Engineering
University of Sydney
Darlington, NSW
Australia

ISBN 978-3-319-43221-2 ISBN 978-3-319-43222-9 (eBook)
DOI 10.1007/978-3-319-43222-9

Library of Congress Control Number: 2016954697

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is aimed at advanced undergraduate and graduate students across a wide range of fields, from computer science and physics to the many current and potential application areas of transfer entropy. Other researchers interested in this new and fast-growing topic will also find it useful, we hope.

It sits at the nexus of information theory and complex systems. The science of complex systems has been steadily growing over the last few decades, with a range of landmark events, such as the formation of the Santa Fe Institute in 1984, and the fundamental work of physics Nobel Laureates Murray Gell-Mann and Phillip Anderson. But precisely defining complex systems proved illusive. There are many examples, properties, ways of simulating and a diversity of theoretical suggestions. But it is only after 30 years that the pieces are finally falling into place.

Information theory, dominated by Claude Shannon's mathematical theory of communication, was one of the great theoretical ideas of the 20th century. It proved a valuable tool in analysing some complex systems, but it was only much later, with Schreiber's transfer entropy, that the relationship between information flow and complexity became apparent.

This book, like any complex system, emerged in parallel, with the synchronisation of ideas and thinking of the four authors. Terry's involvement in information theory goes back a very long way to its use in understanding images and animal vision. But he became interested in complex systems two and a half decades ago and the possibility that information theory would be a key tool was always in the background.

It was through the neuroscience dimension that Terry met Mike, while he was a PhD student at the Centre for the Mind at the University of Sydney. While working there Mike collaborated with David Wolpert of NASA Ames and it was David who introduced Mike to maximum entropy techniques and their application to economic game theory. This collaboration lead to several key findings regarding tipping points in microeconomics, 'persona choice' in behavioural game theory, and contributed significantly to Mike's PhD. During this time Mike also developed the idea of using mutual information as a tool to study financial market crashes in the same way that mutual information had been used to characterise phase transitions in physics.

Terry's collaboration with the University of Sussex began in the mid-1990s, but he and Lionel did not actually engage in any detailed discussions until the Artificial Life Conference in Lisbon in 2007. Lionel, along with Anil Seth, had been working on causality measures, particularly with applications to neuroscience and consciousness, for some while before getting interested in transfer entropy. Lionel then began a series of annual month-long visits to the Centre for Research in Complex Systems at Charles Sturt University, where some of the research in this book had its genesis.

Joe, meanwhile, had been working on transfer entropy during his PhD, finding some extraordinary results for simple systems, such as cellular automata. Although Terry and Joe met in Lisbon, it was not until the IEEE ALife conference in Paris that any sort of real dialogue began. In many ways, that conference was instrumental in formulating the ideas which led to this book.

The structure of the book is a bit like stone fruit, with a soft wrapping of a hard core, although the non-mathematical reader might find it something like climbing a mountain. After a qualitative introduction, Chap. 2 introduces ideas of statistics, which will be familiar to many readers. The going then gets tougher, or at least more mathematical, reaching its zenith in Chap. 4 where the main ideas of transfer entropy are worked out. We adopt Knuth's dangerous bend symbol,  and . The reader already familiar with information theory could perhaps go straight to Chap. 4, but other readers would need the background in Chap. 3. The later chapters of the book introduce a variety of applications, from simple, canonical systems to finance and neuroscience. The full details of Chap. 4 are not necessary to get an idea of the kind of applications covered. Transfer entropy is hard to calculate from real data. Some robust software is now available and new applications are appearing at an increasing rate.

Many people have been influential over the years in the development of this book, and we thank them all. Alan Kragh and John Lewis at Ilford Ltd. gave much encouragement to Terry in the pursuit of theoretical metrics for imaging science. The seminal work by Linfoot and Fellgett was pivotal at that time, although Terry never had the opportunity to meet either. But his real work in information theory began at the Australian National University with Allan Snyder FRS, Mike's PhD supervisor years later. His interest in complexity was stimulated by collaboration with David Green in the 1990s.

Lionel has been supported by the Sackler Centre at the University of Sussex, led by Anil Seth, with whom he has published extensively.

Joe was introduced to complex systems by Terry Dawson, while at Telstra Research Laboratories. This interest was fused with information theory under the guidance of Mikhail Prokopenko, then at CSIRO, now at the University of Sydney. Mikhail played a pivotal role in supervising Joe's PhD, also under Albert Zomaya at Sydney. Joe's work on information theory continued in his postdoc years at the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany, with Juergen Jost.

With regards to this book, Joe thanks in particular Michael Wibral, Juergen Pahle, Greg Ver Steeg and Mikhail Prokopenko for valuable discussions, comments and feedback on draft material.

The authors thank Carolyn Leeder for administrative assistance.

Some of the original research by the authors described in the book was funded by the Australian Research Council.

This book would have taken ten times as long to produce had it not been for Donald Knuth's TEX mathematical typesetting package and Leslie Lamport's LATEX. We use GNUPlot frequently, and Terry uses Emacs extensively almost every day. So thanks, also, to Richard Stallman.

Contents

List of Key Ideas

List of Open Research Questions

List of Key Results

Symbols

β	Inverse temperature
$\mathbf{I}(X:Y \mid Z)$	conditional mutual information
$\mathbf{T}_{Y \to X \mid Z}(t)$	conditional transfer entropy
$\mathbf{H}$	entropy, arbitrary number of dimensions
F	Granger causality
ξ	information discrimination
$\mathbf{t}_{x \to y \mid z}$	conditional local transfer entropy
$\mathbf{i}$	local mutual information
$\mathbf{t}_{x \to y}$	local transfer entropy
$\mathbf{I}$	mutual information between two probability distributions X,Y
$\mathbf{I}(X_1 : X_2 : X_3)$	multi-information among X_i (mutual information for 3 variables)
N_G	number of nodes in a graph or network
Ω	sample space
$\psi(x)$	digamma function
$\rho(x,y)$	Pearson correlation coefficient between x and y
$\eta(x)$	information or surprise. Could also be called local entropy using the definitions of this book
τ	time delay or lag
$\mathbf{T}$	transfer entropy
$\mathbf{A}$	coupling matrix for VAR process
$\mathbf{S}$	VAR process
$\mathbf{G}(p:q)$	cross entropy between p and q
d	embedding dimension
L	path length in a graph

Acronyms

AO	Australian Share Market
CPI	Consumer Price Index
DAX	Frankfurt Stock Index
DDLab	Discrete Dyamics Lab (Andy Wuensche)
DJIA	Dow Jones Share Market
ECA	Elementary Cellular Automata
EEG	Electroencephalography
ET	Effective Transfer Entropy
FTSE	London Stock Exchange (Financial Times Stock Exchange)
G	Cross Entropy, **G**
GDP	Gross Domestic Product
GLM	Generalised Linear Model
GTE	Global Transfer Entropy
JDIT	Java Information Dynamics Toolkit
KLD	$\mathscr{K}(X\|\|Y)$ Kullback–Leibler Divergence between X and Y
kT	Product of Boltzmann's constant k and absolute temperature T
LTCM	Long Term Capital Management
MI	Mutual Information
ML	Maximum Likelihood
QRE	Quantal Response Equilibrium
REA	Relative Explanation Added
ROC	Receiver Operating Characteristic
S&P	Standard and Poor's Stock Index
TB	Trade Balance
TE	Transfer Entropy
TSE	Tononi–Sporns–Edelman (complexity)
XOR	Exclusive OR
XR	Exchange Rate

List of Tables

List of Figures

Chapter 1
Introduction

Decades often acquire an evocative name: the Jazz Age, the Decade of the Brain, maybe the Decade of Moral Hazard for the last decade. Whatever our current decade becomes, this will certainly be the Century of Information. Never in human history has the growth of data been so great. Information comes from cameras everywhere, on street corners, in smart phones, on headbands like a miner's lamp. It comes from supermarket checkouts, credit card transactions, search engines and the vast amount of personal data contributed to social networks.

The Library of Congress (LoC) blog [202] has entertaining illustrations of the vastness of human collected information and its staggering growth. Its book collection is around 15 Terabytes. That's only $1000 or so in disc drives these days. (But most of the LoC's data is on audio and video.) The National Security Agency, in the news in 2013 for the depth and breadth of its surveillance, collects data equal to the LoC *every 6 hours.*

Thus a huge amount of information flows into computer systems, from the NSA to Google, every second of every day. This book is about *information flow.* But rather than be concerned with what the information is, or even where it is going, we are interested here in detecting information flow between systems, from the way they behave and influence one another. Essentially, if we have two systems, things, entities, agents, whatever, for which we can measure some property as a function of time, we want to know, just from this *time series*, if there is information flow from one to the other.

This coupling of time series was studied by Clive Granger, for which he received the Nobel Prize in Economics, although he himself was not an economist [114]. Transfer entropy (TE), the topic of this book, is in many ways a generalisation of Granger causality, discussed in Sect. 4.4.

It can be easier to teach something where the learner has no prior knowledge, than to teach something where there are prior misconceptions. Because we have such exposure to information and to information as a defining characteristic of our lives, we need to step back and build *information theory* from the ground up. The next section makes a start before the full discussion of Chap. 3.

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_1

1.1 Information Theory

In the early decades of the 20th century, Bell Labs laid the foundations for information theory. Hartley introduced the idea of information, and Nyquist the sampling theorem. But it was their protégé Claude Shannon who built a *mathematical theory of communication*, the results of which still stand today [304].

Shannon's interest was how to transmit information over a channel in the most efficient way possible. The channel capacity theorems gave the answer. It transpired that information could even be transmitted perfectly when the channel was imperfect. The analysis introduced the idea of entropy of signals and channels, ideas we take up in Chap. 3.

As a simple example, imagine that a new courier/removal service to meet the insatiable needs of internet shoppers has to equip itself with a set of vans. Obviously, if all the parcels are not the same size, say from packets of tea to beds and sofas, then they would want a set of vans of different sizes. Entropy and coding do something like this for information and messages. Now the courier does not need to know what is in the parcels, just something about the range of sizes in which they occur. So it is with information: the semantics and content are irrelevant.

One of the key ideas introduced by Shannon was *mutual information* (MI), which we shall study in detail in Chap. 3. It describes how much information is shared between two things, or more precisely, sets of things. So we might imagine that the price of coffee beans affects the price of our espresso in the local coffee shop. But, since shops will keep beans in stock, there will be a time lag, and the variations in price of coffee beans and cups of coffee will look most similar if we shift the price of coffee beans forward. We will come back to this example a little later.

1.2 Complex Systems

Complex systems abound in the natural and social world, e.g. across systems as apparently diverse as insect colonies, the brain, the immune system, economies and the world wide web [227]. Yet despite three decades of intense research activity in studying complexity, many big issues remain only partially resolved, including, believe it or not, a good quantitative definition for a complex system. Qualitatively, complex systems are often described as collections of (generally simple) entities, where the global behaviour is a non-trivial result of the local interactions of the individual elements [270]. In attempting to make this description quantitative at the heart of many proto-definitions are the ideas of entropy and information theory in Sect. 1.1, and now, transfer entropy, the theme of this book.

The study of complex systems has benefited enormously from a set of canonical systems, and the book makes extensive use of them for mutual information and transfer entropy too. We take a brief look at each, without any formal definitions or mathematics in the following sections. They are cellular automata (CA) (Sect. 1.2.1), complex networks (Sect. 1.2.4), random Boolean networks (RBN)

(Sect. 1.2.5), spin systems (Sect. 1.2.2) and oscillator populations (Sect. 1.2.3), and flocking behaviour (Sect. 1.2.6). In all cases most of the MI and TE work has been done, with the exception of flocking systems.

1.2.1 Cellular Automata

Of all the canonical systems, CAs are perhaps the most diverse and spread from abstract theory to real-world applications such as modelling soils [120] to traffic [234] and urban sprawl [61]. They have even received the accolade of being the fundamental building block of natural systems [362], a view not universally shared.

CAs appeared in the middle of the last century from the work of Stanislav Ulam and John von Neumann, with theoretical interests in computation itself. From a theoretical point of view several developments stand out: A cellular automaton consists of a set of cells with a finite number of states. These cells are connected together in some way, usually on a rectangular lattice. Each cell has an update rule, usually the same for each cell, and all the cells are updated synchronously or asynchronously at random. It should be clear that there are many sorts of cellular automaton, but our concern in this book will be the simplest, one-dimensional nearest-neighbour lattice types, taken up in Sect. 5.1.

CAs came into the public eye through an article by Martin Gardner in Scientific American [100] describing a two-dimensional cellular automaton, the Game of Life, invented by Princeton mathematician John Conway. The definition is ultra-simple. Each cell can be zero (dead) or one (alive). It has four neighbours, at the major compass points on a rectangular grid. At each step, each cell looks at its neighbours. With one or fewer alive neighbours, it dies. With two or three alive neighbours, it lives (if already alive), but with four, it again dies (overcrowding). If the cell is dead already, it comes to life if there are three and only three live neighbours.

From this simple system, an amazing array of patterns were soon discovered. Interesting examples were: blinkers, which turn on and off; gliders, which move steadily across the grid; and glider guns, which fire an endless stream of gliders. Indeed, many CAs contain such self-organised coherent structures, some propagating, against regular backgrounds (e.g. the "gliders" in the Game of Life) which are conjectured to transfer information from one part of the CA to another. We revisit this conjecture in Sect. 5.1.

Studies of this and other interesting CA systems have led to the following interesting findings:

- CAs are capable of universal computation, in the Turing machine sense [63, 64]. This is not something which concerns us very much here.
- CAs come in four classes: the so-called Wolfram classes [361]:

  I. Fixed point attractor—the CA just stops at one fixed, unchanging pattern
  II. Periodic attractor—the CA cycles through a finite number of states
  III. Chaotic attractor—the CA is chaotic and looks just like noise

IV. Complex—this is the interesting class, to which the Game of Life belongs. These CAs typically contain self-organised coherent structures such as the gliders in the Game of Life.

- CAs exhibit something like a phase transition, first shown by Langton [174] using an informal information-theoretic argument (Sect. 5.1). Most CA researchers feel that there is some sort of analogy to a phase transition, with Langton's λ parameter as the control parameter (Sect. 3.3). But despite several decades of work, there is no completely established way of determining to which class a CA belongs. In fact, one of the authors (Lizier) recently used transfer entropy to revisit the classification of several rules [197] (Sect. 5.1).
- CAs can be reverse engineered; i.e. one can work backwards from the behaviour to the rule, as discovered by Andy Wuensche [366], forming the basis for his software DDLab.

Phase transitions appear frequently within this book, and we can note here that the complex cellular automata rules, class IV, are conjectured to occur at the phase transition. Langton showed that the highest values of *mutual information* occur near the suggested phase transition, something to which we shall return frequently within this book. The Game of Life is not the only such complex automaton, although there are not many complex CAs in comparison with all the others. In fact, a special search procedure is needed to find them, known as maximisation of the input entropy. The details would take us a bit off course, but they are beautifully described in Wuensche's article in Complexity [366].

The treatment of CAs in Chap. 5 is fairly mathematical. The reader interested in the startling visual behaviour of CAs is strongly encouraged to download and explore DDLab.

1.2.2 Spin Models

At first sight, spin models look rather like cellular automata. They are spatial grids with spins, which in the simplest case are binary, but in the Potts variants may have more (integer) states. But the update dynamics are different. There are several update protocols, but the one which will appear most often in this book is *Glauber dynamics* [24].

To understand the dynamics, we need the idea of the *Hamiltonian* of a system, which essentially measures its total energy as a function of some system parameters. In this case the system parameter is the distribution of spins. Two binary spins of the same orientation have lower energy than two of different orientation. The interactions are counted only amongst the nearest neighbours on the grid to compute the Hamiltonian.

The Glauber update is to pick a spin at random and determine the energy change which would result if it flips. Whether or not it flips depends upon the temperature, as discussed quantitatively in Sect. 5.2. The simplest Ising model, which dates back

to Onsager in the 1920s, shows a phase transition as a function of temperature. The model is widely used as a simple theory of magnetic materials, and the phase transition is the change from a material being ferromagnetic to paramagnetic, occurring at the *Curie* temperature.

There are many variants of spin systems: multi-state variants such as the Potts family; and spin glasses, which have more than one spin type, modelling mixtures of materials.[1] But at the time of writing most work in the information theory domain corresponds to the very simple model, which, therefore, will be the primary focus in this book.

1.2.3 Oscillators

Steve Strogatz, known to many for the famous letter to Nature with Duncan Watts on small-world networks [347], has contributed to many aspects of complexity, including a book, *Sync* [317], entirely devoted to synchronisation. Things which vibrate periodically are familiar to us from many domains, notably musical instruments, and in earlier days, clocks which ticked.

One of the earliest observations of synchronisation came from Christiaan Huygens, inventor of the wave theory of light. He observed that an array of clocks on the wall would synchronise. In his day clocks were mechanical systems, which inevitably were not quite perfect, meaning that they all ran at slightly different speeds. But together, coupled by the wall, they become synchronised.

Since Huygens, much has been written about oscillators, from theory and simulation, to applications and observations of all kinds. One quirky example is the way women sharing a house find their periods synchronise. The curious thing is what the coupling is, the equivalent of the wall for Huygens' clocks. It has probably got something to do with human pheromones, but there are strong dissenters from this viewpoint.

Huygens was around in the 17th century, but as with many of these simple systems, new things are still being uncovered in the 21st. Foundations for mathematical study of synchronisation were laid by the Kuramoto model [169, 170], with significant insights gained for example from linear algebra [8, 147]. Chap. 5 takes up the story with new work published by one of the authors of this book (Lizier) showing surprising dynamics of information flow as oscillators synchronise.

1.2.4 Complex Networks

The concept of complex networks has been particularly pervasive in complex systems science [81, 259]. In part, this is due to the centrality of the concept of "local

[1] As an aside, spin glasses had an interesting role in the history of complex systems. They were amongst the first systems to demonstrate *broken ergodicity*.

interactions" between entities (see Sect. 1.2) as giving rise to global behaviour in complex systems. These local interactions can be modelled as *edges* or *links* between pairs of *nodes* (representing the individual entities), giving a graph-theoretic or *network* representation of the system. A pair of nodes may simply have no edge between them, edges in both directions or a directed edge going from only one node to the other. Edges may also be weighted, to give some indication of coupling strength. Networks where coupled pairs are bidirectionally connected (with the same weighting, if applicable) are called undirected networks; and otherwise are directed networks.

The field of complex networks studies the structural properties of such networks. For example, the *degree* of a node is the number of connections it has to other nodes. This can be further specified as *in-degree* and *out-degree* (the number of incoming/outgoing connections) in a directed network. *Path-length* between a pair of nodes refers to the number of edges on the shortest path between them; this may be a weighted sum of edges where weightings exist. The *clustering coefficient* is the proportion of pairs of neighbours of a given node that are also connected by an edge themselves [347]. A *network motif* is a (typically small) sub-graph which is a repeated pattern within a network; e.g. a fully connected triplet of nodes.

The goal of such analysis is to identify common features across various domains, and characterise their functional role. Typically for example, the structures of natural systems (e.g. neural networks, gene regulatory networks) and man-made systems (e.g. power grids) are neither completely regularly structured (like a lattice) nor are they completely randomly connected. Indeed, two very important classes of structures have been identified, and have attracted an enormous amount of attention because they have been found to be incredibly widespread. Watts and Strogatz first described *small-world networks* [347, 346], which balance regular and random network structures to provide both short path length (typically a characteristic of random networks) at the same time as high clustering (typically a characteristic of regular networks). Given the prevalence of these structures in social networks, they provide some explanation for the "six degrees of separation" or "small-world" phenomenon. *Scale-free networks* [19, 20, 21] display a degree distribution where the probability of a node having a given degree is inversely proportional to the degree. Barabási et al. showed that such networks can be constructed via the principle of "preferential attachment" [19, 20]—where new nodes introduced to the system preferentially make connections to nodes in proportion to their existing number of connections, the so-called "rich get richer" phenomenon. A scale-free distribution is highly structured, and is considered to be a signature of self-organised criticality (see [14]). Such networks contain *hubs*—nodes with extremely large degrees which play a key role in the dynamics of the network.

Despite these ground-breaking insights into network structure, the time-series behaviour or dynamics on networks have received less attention and are "much less well understood" [226]. There is a widely-recognised need for fundamental insights into dynamics on networks, and how these are related to the underlying structure [299, 18, 346, 226, 227]. Our next subsection describes one important model of dynamics on networks—random Boolean networks. In the study of dynamics, transfer

entropy has a key role to play, characterising how information is transferred in the local interactions between nodes in an application-independent manner. Chap. 5 will outline how TE is being used to provide insights into the dynamical role of common network structures, with Chap. 7 describing some examples of empirical analysis of dynamics on various networks. Chap. 7 also includes a key application of TE in inferring the structure of complex networks from their time-series dynamics.

1.2.5 Random Boolean Networks

Stuart Kauffman introduced random Boolean networks (RBNs) in 1969 [154] to study gene regulatory networks (GRNs) and initiated rapid growth in theoretical results and diversification of network structure. Some such networks display phase transitions between ordered and chaotic behaviour in the dynamics of their nodes, and thus they come up for consideration in Chap. 5.

An RBN consists of a set of nodes, with two states. In Kauffman's NK model each has exactly K connections to other nodes. Each node has a randomly generated Boolean function which determines whether it will flip in the next step dependent upon its neighbours. In the original model update is synchronous: all nodes are updated simultaneously. Asynchronous update leads to quite different behaviour [135], as is also the case with CAs.

The GRN interpretation of the NK model is that nodes model genes, with the Boolean value of the node modelling gene expression level, connections modelling gene interactions, and the attractors for the network state modelling phenotypes (i.e. different cell types); that is, depending on initial conditions of the RBNs and/or inputs, the same RBN (i.e. GRN) can reach different attractor states (i.e. become different cell types).

Since Kauffman's ground-breaking innovation, RBNs have received a lot of attention. Different node functions have been investigated, such as the simplification of just summing the states of the neighbours. Different connection patterns, reflecting the interest in small-world and scale-free networks, are also of interest and some are discussed further in Chap. 5.

Applications have spread far from biology into the social sciences. In one example, Rivkin uses RBNs to model what makes a successful franchise, arguing that a reasonable level of complexity is required to avoid facile mimicry [286].

1.2.6 Flocking Behaviour

Brighton on the south coast of England has entertainment piers dating back to Victorian times. One of them, the West Pier, has suffered a series of mishaps, from violent storms to major fires. It is now a disused wreck. But it still provides entertainment, courtesy of starlings. These birds congregate in murmurations of tens

of thousands, sometimes from all over Europe, and generate enthralling displays of acrobatic swarms as displayed in Fig. 1.1.

Fig. 1.1 Starlings swarming over Brighton West Pier

Craig Reynolds, who subsequently got hired by the entertainment industry and won an Academy Award, will be found in many books on complex systems for his innovative simulation of bird behaviour: the flocking of boids [282]. The boids model was able to generate realistic simulation of flocking behaviour using only three simple rules for separation, alignment and cohesion between nearby individuals. But flocking goes much deeper than starlings and boids. Vicsek [338] used a flocking model to simulate the phase transition in magnetic systems we discussed in the context of the Ising model (Sect. 1.2.2). In the real biological world, Buhl et al. [51] directly observed a phase transition to collective motion in swarming locusts, with respect to changes in density. The transfer entropy of flocking behaviour as a function of order parameter is an open research question. Couzin [66] interprets criticality in effective flocking behaviour occurring only at intermediate sensory ranges between individual agents in terms of the capacity for information transfer the sensory range allows: too short a sensory range does not allow enough information transfer to form cohesive groups; too large a range permits rampant spreading of irrelevant information which erodes group cohesion. The quantification of information flow leading to synchronisation at a constant order parameter has partially been solved indirectly, since the Vicsek model is equivalent to the Kuramoto oscillator model under certain conditions [59]. We discuss this further in Sect. 5.5.

1.3 Information Flow and Causality

Finally we come to the core of the book: information flow. Even though we adopt the mainstream definition of information from Shannon in this book, there are other definitions, such as Fisher information [274], which has also been linked to phase transitions. When we come to information flow, even within the Shannon framework there is variation, but our focus is, as the title of the book might suggest, *transfer entropy*.

Before we see the detailed mathematics in Chap. 4, we can get an intuitive idea by continuing our coffee bean to cup example. We expect the trend in coffee prices to be reflected later in the trend in coffee shop prices. But we cannot immediately infer from this that the coffee bean price causes the coffee shop price, and we would not expect the price of peanuts to affect the price of coffee. Over the last ten years the coffee bean commodity price has increased by about a factor of three, according to indexmundi.com. In the same time the commodity food and beverage price index has increased by a factor of two (so coffee drinkers should feel aggrieved if they did not invest in coffee beans). The price of peanuts has increased by 2.6, not quite as good an investment as coffee but better than the index.

The problem then is that, because the price of peanuts has been going up relatively fast, it looks as if it might be impacting on the price of coffee. But everything has been going up according to the index. This general inflation creates a sort of common cause effect, which confuses the inference of causality. To get around this we need to extract out such effects. This is what transfer entropy does. It removes from the time-shifted mutual information, the effect of the past price of coffee in the shop, thereby taking out all the general inflation factors in retail food and beverage prices.

There are two distinct features of the transfer entropy approach to information flow in this book:

- The first is that it sits firmly on a huge body of work in economics, for which Clive Granger won the Nobel Prize for his now eponymous causality. Barnett et al. [22] showed some time later that Granger causality and transfer entropy are identical for Gaussian processes. This is taken up in detail in Chap. 4.
- The second is much more philosophical in nature, but of profound significance. We allude in Chap. 3 to the somewhat arbitrary assignment of the term entropy by Shannon. But it turned out to be remarkably prescient. Thermodynamic and Shannon entropy were ultimately reconciled. One might then ask whether there is a link between information and energy. Significant effort went into establishing the thermodynamic cost of computation, starting with work by physicist and Nobel Laureate, Richard Feynman. It took some while to reach consensus, but work by Landauer, Bennett and others [34, 172, 33] ultimately established that computation does not take any energy at all. But the destruction of information during computation *does* cost, at precisely 1 bit per $kT\ln(2)$ Joules of energy, with k being Boltzmann's constant and T absolute temperature. In a 2013 paper

the killer finding by Prokopenko et al. [275, 273] is that information flow, *as measured by transfer entropy,* requires kT per bit of information transferred.

1.4 Applications

The possible applications of transfer entropy ideas are legion, but work to date has mainly been concentrated in neuroscience, with other work in bioinformatics, artificial life, and climate science (Chap. 7), as well as finance and economics (Chap. 6). Two main areas stand out in neuroscience: EEG and neural communication. EEG (electroencephalography) is a technique for recording brain signals from the scalp. The surprising amount of information obtained, with high temporal precision, makes it a widely used, minimally invasive tool. TE serves not only to make links between behaviour and EEG signals but also as a predictor of pathological events such as epileptic fits.

At the neural level, one very difficult question to answer physiologically is whether neural activity in one part of the brain influences or causes activity somewhere else. TE provides a mechanism for quantifying information flow between active neurons.

1.5 Overview

The real work of the book starts in Chap. 2, where the statistical foundations are described. With the statistics in hand, Chap. 3 introduces information theory and Chap. 4 gets to grips with the mathematics of transfer entropy itself.

The first applications appear in Chap. 5, where canonical systems are studied. These systems have been pivotal in developing our theoretical understanding, thus this chapter has a strong theoretical slant too.

The primary application area we discuss is financial markets, taken up in Chap. 6. The remaining applications are spread over numerous fields, including neuroscience, and occupy Chap. 7. The book concludes with some retrospective comments, a discussion of some important open research questions and the exciting opportunities for new applications.

Chapter 2
Statistical Preliminaries

The foundations of information theory are firmly grounded in the field of probability theory; to this end this chapter introduces the technical background needed later. For the purposes of this book we base our notions on those of the frequentist interpretation of probability while acknowledging that this is primarily due to its readiness of exposition. The work of probability theory, certainly as we understand it today, stems from the letters and research of Gerolamo Cardano in the 16th century and Blaise Pascal and Pierre de Fermat in the 17th century. These earliest works centred around probabilities in games of chance such as cards and dice as well as theological issues such as Pascal's wager (how you should bet your eternal soul based on the probable existence of God [123]). These ideas were later formally axiomatised by Kolmogorov [164], and the rigorous foundations of the field are now well established. There still remain strong differences in the philosophical foundations of probability theory, but these are of no significance in this work, although interesting historical notes are mentioned as they arise. Andrey Kolmogorov was himself one of the greatest probabilists of all time, and he noted: "The epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity." [163]. In a similar flavour to that of the earliest probability theorists, Sting, one of the greatest lyricists of recent times, wrote for the song *Shape of my Heart*:

> He deals the cards as a meditation
> And those he plays never suspect
> He doesn't play for the money he wins
> He don't play for respect
>
> He deals the cards to find the answer
> The sacred geometry of chance
> The hidden laws of a probable outcome
> The numbers lead a dance

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_2

2.1 Set Theory

We want to formalise the (probabilistic) relationships between individual elements of quite different systems. To do this we begin with an individual element in a set, and we label each element ω_i where $i \in \{1,2,\ldots,M\}$. Each possible state that $\omega \in \{\omega_i\}_M$ can take represents a sample point in a set of sample points called the sample space: $\Omega = \{\omega_1,...,\omega_M\}$ and we define the size (*cardinality* or the number of elements in the set) of the sample space as $|\Omega| = M$. For example for a coin toss we have $\Omega_{\text{coin}} = \{\text{heads},\text{tails}\}$, for a six-sided dice $\Omega_{\text{dice}} = \{1,2,3,4,5,6\}$ or for a pack of 52 playing cards the suits $\Omega_{\text{suits}} = \{\heartsuit,\diamondsuit,\spadesuit,\clubsuit\}$, the face cards $\Omega_{\text{face}} = \{\text{King},\text{Queen},\text{Jack}\}$ or the non-face cards $\Omega_{\text{nonface}} = \{10,9,8,7,6,5,4,3,2,\text{Ace}\}$.

Each of these examples has in common the fact that each event in each Ω has equal probability of occurring: drawing any face card from Ω_{face} has a probability of $\frac{1}{3}$, and tossing a coin and getting either a head or a tail is $\frac{1}{2}$ each. We can go from the frequency with which an event occurs to the probability of occurrence by simply taking the frequency of a single event in Ω and dividing it by the total number of *trials* N that have taken place. For example a coin tossed 1000 times might come heads 494 times, and the estimate of the probability of the event Heads is $p_{\text{est}}(\text{Heads}) = \frac{\text{freq(Heads)}}{\text{Trials}} = \frac{494}{1000} = 0.494$. As the total number of events increases so too does the accuracy of the estimate of the probability of each event. This is called the *frequentist interpretation* of statistics, and it is only one of several different interpretations (see Section 2.3.3 and [146] for Bayes' Theorem, the foundation of an alternative interpretation of statistics). The elements ω_i are called elementary events; they are the indivisible elements of the sample space Ω.

But we also want to define subsets of Ω that are larger than elementary events. For example if $\Omega_{52\text{ cards}} = \{\text{A pack of 52 playing cards}\}$ then $\omega_i = \{\text{a unique playing card defined by suit and value}\}$, but instead of the set of single playing cards we may be interested in the set of cards whose suit is spades. Clearly this set is smaller than $\Omega_{52\text{ cards}}$ but larger than any ω_i, so we define the subsets of Ω: A, B, C etc. so that for a set A: $A \subseteq \Omega$. So a collection of ω_i is a subset of Ω and they are denoted A, B etc., and the cardinality of these sets is denoted $|A|$, $|B|$ etc. If A is a subset of B, we write $A \subset B$ and if both sets contain the same ω_i then $A = B$. The union of two sets $A,B \subset \Omega$ is written $A \cup B = \{\omega \in \Omega \,|\, \omega \in A \text{ or } \omega \in B\}$, the intersection is written $A \cap B = \{\omega \in \Omega \,|\, \omega \in A \text{ and } \omega \in B\}$, and two sets are disjoint if their intersection is the empty set: $A \cap B = \emptyset$.

For example a pack of 52 cards with no jokers can be divided into non-overlapping subsets of $\Omega_{52\text{ cards}}$ in distinct ways, for example four suits: $A_\heartsuit$, $A_\diamondsuit$, $A_\spadesuit$, $A_\clubsuit$ or face and non-face cards: B_{face}, B_{nonface}. Each set is a subset of $\Omega_{52\text{ cards}} = A_\heartsuit \cup A_\diamondsuit \cup A_\spadesuit \cup A_\clubsuit = B_{\text{face}} \cup B_{\text{nonface}}$. In a set of English or French playing cards $A_\heartsuit \cap B_{\text{face}} = \{\heartsuit\text{King}, \heartsuit\text{Queen}, \heartsuit\text{Jack}\}$.

2.2 Discrete Probabilities

With these notions of sets we can define probability spaces. A probability space is defined by the triple $\{\Omega, X, p\}$ where Ω is the space of all elementary events, X is a set of disjoint subsets of Ω called events (whose union is Ω) and p is a probability function that maps events in X to the closed unit interval $p : X \rightarrow [0,1]$. We also define a random variable $x : \Omega \rightarrow X$ that maps from the sample space to events.

We can now define the probability p of a discrete random variable x belonging to the event $x_i \in X$ as the result of a statistical process in the following way: $p(\{\omega \in \Omega : x(\omega) \in x_i\}) = \frac{|x_i|}{|\Omega|}$ assuming there is a uniform probability of any elementary event ω occurring. For example we want to know what probability of drawing a diamond face card from a pack of 52 cards if the probability of choosing any one of the 52 cards is the same as any other (uniform distribution) $x : \omega \in \{\diamondsuit\text{Jack}, \diamondsuit\text{Queen}, \diamondsuit\text{King}\} \rightarrow x_1 = \{\diamondsuit\text{face cards}\}$, $|x_1| = 3$, $|\Omega_{52\,\text{cards}}| = 52$, $p(\omega \in \Omega : x(\omega) \in x_1) = \frac{3}{52}$, for which we might more simply write $p(x = x_1)$ or $p(x_1)$ if there is no confusion, and we refer to the distribution of p over $x \in X$, $p(x)$ as a *probability distribution function* (PDF). While the cardinality of sets is a useful measure over equally likely outcomes of elementary events, non-uniform probabilities over elementary events are also possible.

For a probability space $\{\Omega, X, p\}$ for which $|X| = M_X$ is at most countably infinite we have the following three axioms:

1. $p(\Omega) = 1$
2. $\sum_{i=1}^{M_X} p(x_i) = 1$
3. $p(x_i) \geq 0 \ \forall$ disjoint events $x_i \subset X$

By disjoint events we mean $\bigcup_{i=1}^{N} x_i = X$, $x_i \cap x_j = \emptyset \ \forall \ i \neq j$, and so we can write $p(x_i \cap x_j) = p(\emptyset) = 0$, $p(x_i \cup x_j) = p(x_i) + p(x_j)$ and consequently

$$P(\Omega) = P(X) = 1$$

For $A, B \in X$ (not necessarily disjoint) we have the following natural relations:

1. $A = \Omega \ \Rightarrow \ p(A) = 1$
2. $A = B \ \Rightarrow \ p(A) = p(B)$
3. $A \subset \Omega \ \Rightarrow \ p(A) < 1$
4. $A \subset B \ \Rightarrow \ p(A) < p(B)$

An illustrative example is the tossing of a fair coin. In this case $\Omega = \{\text{H}, \text{T}\}$, and possible x_i sets are: $x_i \in \{\emptyset, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{T}\}\}$. The probability function is then $p(\{\text{H}\}) = p(\{\text{T}\}) = 0.5$, $p(\emptyset) = 0$ and $p(\{\text{H}, \text{T}\}) = 1$, where the last two expressions are read: "The probability of neither heads nor tails is zero" and "The probability of either heads or tails is one", respectively.

2.3 Conditional, Independent and Joint Probabilities

We want to extend these ideas to multiple and joint events, and the probabilistic relationships between them.

2.3.1 Conditional Probabilities

In order to do this we consider the probability space $\{\Omega, X, p\}$ and the probability that the outcome of a process is an event x_i that is the intersection of two sets A, B so that $A, B, x_i \in X$, $p(\omega \in A \cap B : x(\omega) \rightarrow x_i) \in [0,1]$ or simply $p(A \cap B)$. As $A \cap B \subseteq A$ and $A \cap B \subseteq B$ so $p(A \cap B) \leq p(A)$ and $p(A \cap B) \leq p(B)$ (relation 4 above), and we define the conditional probability of A given B as

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \in [0,1]. \tag{2.1}$$

We read this as: "The probability of A given that we know we are in state B" or more simply: "The probability of A given B". Note that $0 \leq p(A \cap B) \leq p(B) \leq 1$ tells us that $p(A|B) \in [0,1]$.

2.3.2 Independent Probabilities

There is also another special case that is important to the work that will follow later: independence of two random processes can be demonstrated through the conditional probabilities. Two probabilities are considered independent of one another iff (if and only if) their joint probability is equal to the product of their individual probabilities:

$$p(A \cap B) = p(A)p(B). \tag{2.2}$$

This relationship can be demonstrated through the use of the conditional probabilities:

$$p(A \cap B) = p(A)p(B) \tag{2.3}$$

$$\iff p(A) = \frac{p(A \cap B)}{p(B)} = p(A|B), \tag{2.4}$$

and it also follows directly that $p(B) = p(B|A)$ so the last line tells us that knowing B changes nothing regarding the probability of A, i.e. A and B do not depend on one another and are therefore statistically independent of each other.

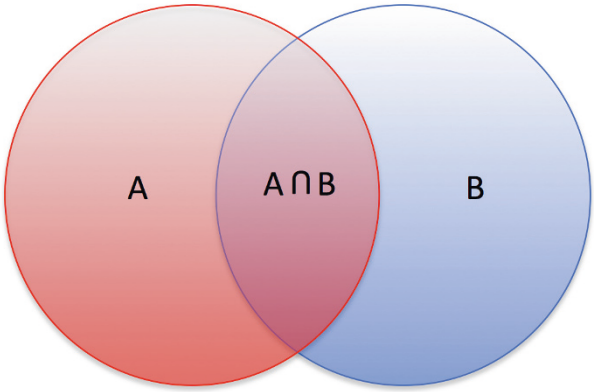


Fig. 2.1 Taking the areas in this Venn diagram as representing the relative occurrence of the events in sets A, B and $A \cap B$, then $p(A \cap B) = \frac{area(A \cap B)}{area(A \cup B)} = \frac{area(A \cap B)}{area(A)+area(B)-area(A \cap B)}$, $p(A|B) = \frac{area(A \cap B)}{area(B)}$ and $p(B|A) = \frac{area(A \cap B)}{area(A)}$.

2.3.3 Joint Probabilities

Now we wish to extend our probabilities of single events to probabilities of joint events given two or more random processes. In the simplest case we have two processes, but this can be extended directly, so we have two probability spaces: $\{\Omega_x, X, p_x\}$ and $\{\Omega_y, Y, p_y\}$, two random variables x and y and a joint probability space $\{\Omega = \Omega_x \times \Omega_y, XY, p\}$. If a joint event $(x_i, y_j) \in XY$ is formed by the co-occurrence of event $x_i \in X$ and event $y_j \in Y$ then the joint probability is given by $p(x = x_i, y = y_j) = p(x(\omega_x) \in x_i, y(\omega_y) \in y_j) \in [0,1]$ or more simply $p(x_i, y_j)$. Note that summing over all joint events equals unity: $\sum_{i,j} p(x_i, y_j) = 1$. The following relations are often useful (see Fig. 2.1):

$$p(x_i) = \sum_j p(x_i, y_j) \text{ (called marginalisation)}, \tag{2.5}$$

$$p(x_i, y_j) = p(x_i|y_j)p(y_j). \tag{2.6}$$

This last relation leads directly to a very useful result called Bayes' theorem:

Theorem 2.1. *Given a joint probability distribution $p(x,y)$ and the related marginal distributions $p(x)$ and $p(y)$, Bayes' theorem states that:*

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{2.7}$$

This theorem allows us to translate a conditional probability of x given y to that of y given x, an exceptionally useful result for many applications in statistics and the applied sciences [179].

In light of today's social climate in which science is sometimes seen as oppositional to religion, Thomas Bayes [b. circa 1701, d. 1761], for which Eqn. 2.7 is named, is a striking character. He was a mathematician and a Presbyterian minister who is known to have published only two works, one a religious work entitled *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* and the other a mathematical work entitled *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst*. Edwin Thompson Jaynes wrote extensively on Bayesian statistics (a form of statistics founded on Bayes' work and totally different in principle from the Kolmogorov statistics used in this book) and was drawn to reach the following conclusion: "What we consider to be fully half of probability theory as it is needed in current applications [...] is not present at all in the Kolmogorov system. Yet [...] we find ourselves, to our own surprise, in agreement with Kolmogorov and in disagreement with his critics, on nearly all technical issues. [...] Each of his axioms turns out to be, for all practical purposes, derivable from [a set of] desiderata of rationality and consistency." ([146], Preface, page xxi, 2009 edition). These *desiderata* of rationality and consistency are founded upon an 18th century cleric's work on probability, thereby laying the foundations for the extension of rational logic to the use of statistics in empirical scientific enquiry in full use today.

2.3.4 Conditional Independence

A special type of independence occurs when two random variables a and b are not independent of each other in that $p(a,b) \neq p(a)p(b)$ but instead are indirectly related to one another via a third random variable c in the following fashion:

$$p(a,b|c) = p(a|c)p(b|c). \tag{2.8}$$

This is called *conditional independence* because a and b are independent of one another once the dependency upon c is accounted for. Such relationships can be further generalised to an arbitrary number of random variables. To do so we take two different sets of possible outcomes called $A_i \subseteq \{a_1,\ldots,a_n\}$ where all a_i are random variables. The following definition generalises the conditional independence of Eqn. 2.8:

$$p(a_1,a_2,\ldots,a_n) = \prod_k p(a_k|A_k), \tag{2.9}$$

such that $A_k \cap a_k = \emptyset$. Such probability separation is the basis of Bayesian networks. For example if we had the following joint probability:

$$p(a_1,a_2,a_3,a_4,a_5) = p(a_5|a_4,a_3)p(a_4|a_2)p(a_3|a_2)p(a_2|a_1)p(a_1), \tag{2.10}$$

it can be described using the network structure shown in Fig. 2.2. If we call the random variables a_i *nodes* then the sets of nodes A_k upon which each a_k is conditionally dependent are called the *parent nodes* of node a_k. An important example of these networks are the so-called Chow–Liu trees that are formed by connecting together nodes such that each node is conditionally dependent on only one other by maximising the Kullback–Leibler measure (see Chap. 3 (Sect. 3.2.4) for definitions) between the original probability $p(a_1,\ldots,a_n)$ and an approximation to $p(a_1,\ldots,a_n)$ given by $\prod_k p(a_k|A_k)$, where $A_k \cap a_k = \emptyset$ and A_k is a single-element set.

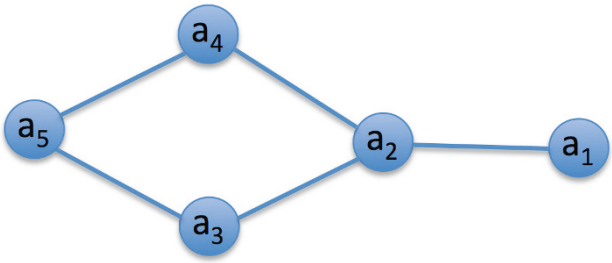


Fig. 2.2 A network of statistical dependencies between the stochastic variables a_i

2.3.5 Time-Series Data and Embedding Dimensions

A time series is a temporally indexed sequence of data points or events; the index is usually denoted by t and can be either be a continuous parameter: $t \in \mathbb{R}$; or a discrete parameter: $t \in \{0,1,2,3,\ldots\}$. So in general, if there is a sequence of temporally ordered random events, we denote the random variable $x_{t_i} \in \{x_{t_1}, x_{t_2}, x_{t_3}, \ldots\}$ where t_i is either discrete $t_1 = 1$, $t_2 = 2$, $t_3 = 3$ etc. or continuous $t_i \in \mathbb{R}$. For example we may be interested in the arrival of customers at a checkout queue at the supermarket; this is a stochastic arrival process that is continuous in time where the arrival of the i^{th} person at the back of the queue occurs at time $t_i \in \mathbb{R}$, $i \in \{1,2,\ldots\}$. This process can be viewed in two distinct ways: we might assume that the number of arrivals in the time interval $[0,T]$ is a Poisson distribution (see Sect. 2.5.2) with mean arrival rate of λt, or equivalently the time between arrivals (the inter-arrival times) $t_{i+1} - t_i$ is a continuous exponentially distributed stochastic process with mean inter-arrival time of λ^{-1}.

An important aspect of time series analysis is how we can infer the underlying system dynamics from a single observation of a time series, rather than many example time series of the same system in which case we could study the statistics of the time series directly. So we consider the *embedding dimension* of a time series of data points. One of the most difficult problems we are faced with when looking at a system with unknown and possibly chaotic dynamics is how to reliably reconstruct

the system dynamics from a single time series of sampled data points. Takens [320] was able to prove that we can reduce the amount of data we need to sample from a d-dimensional system by sampling a single time series.

To illustrate this *Takens* or *time-delay embedding*, let the value of y_t be a dependent variable evaluated (or observed) at a discrete time point $t \in \{1,2,\ldots\}$. Assume that y_t is a function of a d-dimensional state space $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^d\}$ in the following way:[1]

$$y_t = f(\mathbf{x}_t). \tag{2.11}$$

The practical goal of dynamical systems analysis is often to try and accurately recreate the state-space dependency of y_t. However we very often do not know the *complete* state space $\mathbf{x}_t$ and trying to measure it and then explicitly construct Eqn. 2.11 can be an impractical or even impossible task. Fortunately we can write an alternative functional form for the *state* $\tilde{y}_t$ which instead of being dependent on $\mathbf{x}_t$ is dependent on the past history of y_t alone:

$$\tilde{y}_t = \{y_{t-\mathbf{m}\tau}, y_{t-(\mathbf{m}-1)\tau}, \cdots, y_{t-1}\}, \tag{2.12}$$

where τ is called the embedding delay or lag time (the time between successive observations of y_t) and $\mathbf{m}$ is the total number of past data points in the delay vector. Note that $\tilde{y}_t \neq y_t$, but they are very closely related to one another and importantly the same non-linear dynamics, such as chaotic motion, that are observed in y_t are preserved in $\tilde{y}_t$. In this formulation $\tilde{y}_t$ has as its *state dependency* on the $\mathbf{m}$ past observations, so $\{y_{t-\mathbf{m}\tau}, y_{t-(\mathbf{m}-1)\tau}, \ldots, y_{t-1}\}$ is quite literally an alternative (re)construction of the state space dependency of y_t. Takens' main state space reconstruction result then says:

Key Idea 1: *We can accurately reconstruct the* state *of a d-dimensional, non-linear dynamical system $y_t = f(\mathbf{x}_t)$ by observing the $\boldsymbol{m} : d \leq \boldsymbol{m} \leq 2d+1$ past data points of the one-dimensional time series y_t.*

In this formulation $\mathbf{m}$ is called the embedding dimension, and there are variations of this scheme in which τ is not a constant but instead is allowed to vary (e.g. [85]), but this will not be important in the work that follows.

2.3.6 Conditional Independence and Markov Processes

An important type of system is one in which the current state, or more generally a finite number of previous states, influences the outcome of the next event in a temporal sequence of statistical outcomes. This is not (usually!) the case for a series

[1] y_t should be coupled to each dimension of $\mathbf{x}_t$ for the following to hold.

of dice throws: if I throw a 5 now, the fact that it is a 5 has no effect on what value the dice will have when I throw it next, i.e. the throw of a dice is independent of the outcomes of all previous throws; such processes are called *memoryless processes*.

In some systems it is important to consider, at the very least, the current state of the system, as it will influence what state the outcome of the next event will be. For example, if you are gambling in Las Vegas and you start with a gambling purse of \$1000, then the state of this purse after each gamble depends on how much you have before that gamble. Note that it only depends on the state immediately prior to the next gamble, not on the contents of the purse before any previous gamble; the current contents of the gambling purse is said to be a Markov process [99] of order 1 (or *memory* 1). We can describe the state S_t of the purse after a \$10 bet is placed on the t^{th} toss of a fair coin as

$$p(S_t|S_{t-1},S_{t-2},\ldots,S_1) = p(S_t|S_{t-1}) \tag{2.13}$$

$$= \frac{p(S_t,S_{t-1})}{p(S_{t-1})}, \tag{2.14}$$

because there is a 50% chance that $S_t = S_{t-1} + \$10$ if the gamble pays off and a 50% chance $S_t = S_{t-1} - \$10$ if the gamble does not pay off. The second relationship, Eqn. 2.14, follows directly from Eqn. 2.6. We could make this a memoryless process (a Markov process of order 0) by considering only the changes in the value of the gambling purse; then, providing there is \$10 in the purse, its value goes up or down by \$10 with a probability of 50% and this change in purse value is independent of the previous purse value. So the following definition is very useful:

Definition 2.1. *A Markov process of order m is a stochastic, time-ordered process for which the following relationship holds:*

$$p(S_t|S_{t-1},\ldots,S_m,\ldots) = p(S_t|S_{t-1},\ldots,S_m); \tag{2.15}$$

i.e. it is conditionally independent of its past given the previous m states. There is an important assumption of a Markov process: the statistical process that generates the data from which the probabilities are calculated must be time invariant; such a process is called *stationary*. Time invariant means that for an order n Markov process $P(S_t|S_{t-1},\ldots,S_{t-n}) = P(S_{t'}|S_{t'-1},\ldots,S_{t'-n})\ \forall t$ and t'. In the above example of a gambling purse in Las Vegas: if the probability of the coin coming up heads or tails does not change over time then the statistical process (the tossing of the coin) is stationary. If this property does not (at least approximately) hold it becomes difficult to draw reliable statistical conclusions about the system as the relationship between past and future outcomes is different at different times. This is true for all of the systems we will consider in this work. Data is often collected over time and the temporal dependencies between current and past values of a statistical variable as well as statistical dependencies between different variables will be important, so the statistical relationships need to be tested for stationarity so that we can be confident that the conclusions we draw are reliable.

Note the distinctions between the last two definitions: conditional independence that reflects independence between stochastic variables given a suitably complete set of other statistical variables are accounted for, whereas a Markov process is conditionally independent of its past once a sufficiently complete set of historical outcomes have been accounted for.

2.3.7 Vector Autoregression

In statistics, vector autoregression (usually abbreviated to VAR) methods refer to a class of models whose goal is to understand the *linear* relationships between multiple statistical processes (see for example Campbell et al.). In this case we have a vector of n different, potentially coupled, statistical processes that generate a sequence of data points over time; the state of the system at time t is given by $\mathbf{S}_t = [S_t^1, S_t^2, \ldots, S_t^n]$. Then the following linear system of equations is used as a model of the relationships between the different processes:

$$\mathbf{S}_{t+1} = \mathbf{A}\mathbf{S}_t + \boldsymbol{\varepsilon}_{t+1}. \tag{2.16}$$

Here $\mathbf{A}$ is an $n \times n$ matrix of coupling coefficients between the different processes and $\boldsymbol{\varepsilon}_{t+1}$ is a vector of statistical perturbations each process experiences between t and $t+1$. The goal of VAR-type analysis is to estimate the matrix $\mathbf{A}$. Eqn. 2.16 is explicitly a lag-1 process, sometimes denoted *VAR(1)*, as only the previous state of the vector, $\mathbf{S}_t$, is used to estimate $\mathbf{S}_{t+1}$; this can be generalised to arbitrary lags, but the notation can get somewhat cumbersome. For a two-process, lag-1 VAR process, Eqn. 2.16 reduces to

$$S_{t+1}^1 = A^{1,1}S_t^1 + A^{1,2}S_t^2 + \varepsilon_{t+1}^1, \tag{2.17}$$

$$S_{t+1}^2 = A^{2,1}S_t^1 + A^{2,2}S_t^2 + \varepsilon_{t+1}^2. \tag{2.18}$$

Note that the following relationship for the stochastic variation terms is assumed to hold: $E\{\varepsilon_t^i\} = 0$, $\mathbf{c}(\varepsilon_t^i, \varepsilon_{t-2}^i) = 0$, we will cover the definitions of expectations ($E\{x\}$) and covariance ($\mathbf{c}(x,y)$) next.

2.4 Statistical Expectations, Moments and Correlations

One of the most useful purposes to which probabilities are put is in mathematical expectations. For a given numerical quantity $A(x_i) \in \mathbb{R}$ that can be ascribed to the outcome of a statistical process (recall that not all outcomes are naturally numerical) and for which a probability $p(x_i)$ can be defined for all $x_i \in X$, we can further define the expected value of $A(x_i)$:

$$E\{A(x)\} = \sum_{x_i \in X} p(x_i)A(x_i). \tag{2.19}$$

This is also called the *mean* or *first moment* of the random variable x. The $A(x_i)$ can be quite general in form, providing that the x_i is an event in a stochastic process. A very important example is the variance of a stochastic process, defined as

$$E\left\{(x - E\{x\})^2\right\} = \sigma(x)^2, \tag{2.20}$$

where $\sigma(x)$ is called the *standard deviation* of the variable x and $\sigma(x)^2$ is called the variance, often simply denoted var(x). For multivariate processes the expectation is a direct extension of Eqn. 2.19:

$$E\{B(x,y)\} = \sum_{(x_i,y_j) \in XY} p(x_i,y_j)B(x_i,y_j) \tag{2.21}$$

for any $B : XY \rightarrow \mathbb{R}$ with XY the joint event space as defined above; we may sometimes write more generically $E\{B(x,y)\} = E\{x,y\}$. Note that in general $E\{x,y\} \neq E\{x\}E\{y\}$, i.e. the joint expectation of random variables x and y is not equal to the expectation derived by assuming that x and y are independent random processes. This motivates the definition of the *Pearson correlation coefficient*, a measure of the degree to which two random processes diverge from independence:

$$\rho(x,y) = \frac{E\{x,y\} - E\{x\}E\{y\}}{\sigma(x)\sigma(y)} \in [-1,1]. \tag{2.22}$$

Note that $\rho(x,y)$ is a measure of *linear dependence*: if x and y are related to one another by the linear relationship $y = ax + b + \varepsilon$ where ε represents some unexplained statistical variation in the relationship between the two random variables, then $\varepsilon = 0$ and $a \neq 0$ implies a perfect linear relationship between x and y, i.e. one in which there is no unexplained variation *between the two variables*, even if x itself has some unexplained statistical variation. In this case $\rho(x,y)$ is either 1 or -1 depending on the sign of a. For a non-linear relationship between x and y, $\rho(x,y)$ will not pick up all of the covariation between the variables. In the case of $E\{x,y\} = E\{x\}E\{y\}$ then $\rho(x,y) = 0$ and the two processes are considered to be (linearly) independent, even if there is a non-linear relationship between x and y.

Note that, in the definition of the variance given by Eqn. 2.20, we can extend the definition to the variance between two stochastic variables:

$$E\{(x - E\{x\})(y - E\{y\})\} = E\{x,y\} - E\{x\}E\{y\} = \Sigma(x,y). \tag{2.23}$$

This is the covariance between two variables and is used to measure the degree to which one stochastic variable linearly varies with another (note that one variable does not cause the other to vary; they both vary together, and no causation in either direction is implied). By comparing Eqn. 2.23 with Eqn. 2.22 we see that the Pearson correlation coefficient is simply a regularised form of the covariance: di-

viding the covariance by the product of the two variances bounds the covariance to lie between [–1, 1]. Without this regularisation it is very difficult to compare the covariance of two different systems because the variances reflect the different scales of different systems. For example imagine comparing the covariance of height with weight of a population of people with the covariance of income with education level for the same population: the two covariances would be vastly different and tell you nothing about the similarity of the statistical relationship between (height vs. weight) and (income vs. education). In this way the Pearson correlation coefficient allows us to systematically compare the statistically covarying nature of quite different data sets.

2.5 Probability Distributions

Perhaps the three most significant distributions are the binomial distribution, which is used to analyse two-outcome statistical processes, the Poisson distribution frequently used to model independent arrival times and the Gaussian (or normal) distribution, important due to its deep connection with many of the most important statistical proofs and many empirical research problems. These three distributions are also connected to one another, as the binomial distribution can be thought of as a discrete approximation to the Gaussian distribution under certain circumstances and two independent Poisson processes are related to the binomial distribution by a conditional probability distribution of these two processes.

2.5.1 Binomial Distribution

The binomial distribution deals with perhaps the simplest possible statistical process we are ever likely to be interested in: a sequence of n statistical experiments of a two-outcome process where one outcome has probability p and the alternative has probability $1-p$. An obvious example of such a situation is the tossing of either a fair ($p = 0.5$) or biased ($p \neq 0.5$) coin a total of n times. A more sophisticated example is the daily returns of an equity traded on a share market over an n-day period. This second example implicitly underpins a great deal of modern risk in finance [140].

Let us say we have a biased coin where our random variable for the t^{th} toss of the coin x_t has probability $p(x_t = \text{Heads}) = p$ and $p(x_t = \text{Tails}) = q$. Then because the tossing of a coin is a memoryless process (each toss is independent of all previous tosses), the probability of n tosses resulting in k Heads and $n-k$ Tails is $\binom{n}{k}\prod_{i=1}^{n} p(x_i) = \binom{n}{k} p^k q^{n-k}$. The factor $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ comes from the different number of ways in which the sequence of coin tosses can result in k Heads and $n-k$ Tails. This factor is a minimum for $k = 0$ or $k = n$ where there is only 1 possible way

to throw all Heads or all Tails from n tosses (by convention $0! = 1$) and maximised for $k = n-k = n/2$ (even number of tosses) or $k = n/2+0.5$ (odd number of tosses).

For a fair coin with $p = q = 0.5$, both outcomes have the same probabilistic weighting and the binomial distribution reduces to the counting of the different possible ways in which n tosses can result in k Heads and $n-k$ Tails. This is exactly the result for the (normalised) counting of the cardinality of sets described in Sect. 2.1. The mean of the binomial distribution is simply the expectation of the outcome after n trials: $\langle x \rangle_n = np$, and the standard deviation is $\sigma(x) = \sqrt{npq}$.

2.5.2 Poisson Distribution

The Poisson distribution was first introduced by Siméon Poisson in his work on criminal and civil matters of law. Poisson was interested in modelling the probability of discrete events occurring within a certain interval of time. In order to model this process (the Poisson process) he proposed the following probability distribution for a random variable x representing the number of arrivals per unit time:

$$p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2.24}$$

where λ is the only free parameter, $e = 2.71828$ is the base of the natural logarithm and k is the number of events that were observed to have occurred within the given time interval.

Suppose now that you have an office during which students can come by and discuss the lectures and course materials with you. Through past experience you know that your students arrive at approximately 4 per hour $= \lambda$ (sometimes called the arrival intensity), but you have not had lunch yet and you want to duck out for half an hour, grab a bite to eat and come back. What is the probability that during this time at least one student will arrive and you will not be there? The probability that no student arrives is $p(x = 0)$, and so the probability that more than one arrives is simply $1 - p(x = 0)$, where $p(x) = \frac{\lambda^k e^{-\lambda}}{k!}$ with $\lambda = 0.5 * 4 = 2$, $k = 0$ and $0! = 1$ so $p(x = 0) = \frac{2^0 e^{-2}}{0!} = e^{-2} = 0.135$ so the probability of at least one student going away disappointed is 86.5%.

Note that this is an important example of the need for the arrival process to be a *stationary* probability distribution: the arrival intensity λ needs to remain fixed over time otherwise, if the arrival intensity changes over time, your estimate will also be incorrect.

2.5.3 Continuous Probabilities

Before moving on to the continuous Gaussian distribution, we need to consider some important ideas regarding the relationship between continuous and discrete probability distributions. In a continuous random process, a random variable x can take any value in a continuous real-valued space. For simplicity we only consider $x_i(\omega) \in \Omega \subseteq \mathbb{R}$, i.e. events x_i that are strictly elementary events. The cardinality of Ω is no longer countably infinite; if $-\infty \leq \Omega \leq \infty$ we say the *support* of x is the real line, as is the case for the Gaussian distribution (see Sect. 2.5.4).

Next we consider a sample space Ω that has a total order, this allows us to place individual events, either elementary or composite, in a ranked order with respect to each other, a technique that will be very useful in discretising continuous sample spaces. For every $a,b,c \in \Omega$ there is a relation $\preceq$ (note: different from $\leq$, see below) for which the following properties hold:

1. $a \preceq a$ (reflexivity)
2. if $a \preceq b$ and $b \preceq a$ then $a = b$ (anti-symmetry)
3. if $a \preceq b$ and $b \preceq c$ then $a \preceq c$ (transitivity)
4. $a \preceq b$ or $b \preceq a$ (totality)

We note two important examples of a total order: the set of real numbers and the integers, both ordered with the usual binary relation of "is less than or equal to": $\leq$. We can readily construct subsets of a totally ordered set that are also totally ordered sets with the same relation $\preceq$ as the original set. For example consider a variable $x \in \mathbb{R}$ such as the height of a person in metres (m); this is a totally ordered set for which we can define subsets that are also totally ordered, for example $\{1.65\text{ m} \leq x \leq 1.85\text{ m}\}$ is less than $\{1.95\text{ m} \leq x \leq 2.05\text{ m}\}$.

Now if x is a random variable, we can extend our previous definitions to a continuous space X using the same notation from Sect. 2.2. For $a,b,\omega \in \Omega$, we define an event as a *subset* of the sample space $x_i = \{\omega \,|\, a \leq x(\omega) \leq b\} = \{a \leq x \leq b\}$, and we define a probability p just as we did before: $p(\{\omega \in \Omega : x(\omega) \in x_i\}) \in [0,1]$. We can also look at this in terms of a continuous probability function, explicitly:

$$p(\{\omega \in \Omega : x(\omega) \in x_i\}) \equiv \int_a^b p(x)dx, \tag{2.25}$$

where $p(x)$ for continuous x is the *probability density function* (PDF)[2] at x. Analogous to axioms 1 and 2 in Sect. 2.2 we have:

$$\int_\Omega p(x)dx = 1, \tag{2.26}$$

and $p(x) \geq 0$ in analogy to axiom 3 in Sect. 2.2.

[2] The attentive reader will notice that we use PDF as an abbreviation for both probability distribution function and probability density function; one can decipher which it refers to by whether the argument is discrete or continuous.

For example if Ω were all males living in Australia and our random variable of interest was the heights of all males living in Australia, then $x : \{\text{Heights of Australian males}\} \rightarrow x_i \subset \mathbb{R} = \text{height in metres}$. Naturally, the height of an individual is a continuous value, and because individual outcomes that are continuous are unique (no two people ever have *exactly* the same height), constructing probabilities over intervals of continuous outcomes allows us to discretise the outcomes (in the case of heights this is the same as rounding our data to the nearest centimetre for example) and allows us to say there were y Australian males with height of z metres (to the nearest cm).

2.5.4 Gaussian Distribution

Often described as "The Prince of Mathematics" and the "greatest mathematician since antiquity", Carl Friedrich Gauss' [1777–1855] personal motto *pauca sed matura* (few, but ripe) is rather ironic in light of the very prolific career of this quintessential polymath for whom there are over 100 eponyms. The distribution for which he is most famous is called the Gaussian distribution, and it is best introduced in its functional form:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in [-\infty, \infty], \tag{2.27}$$

where the mean is μ and the standard deviation is σ. We can construct a discrete distribution from the Gaussian by noting that x is totally ordered, defining

$$p(\Delta_i x) = p(x'_i \leq x < x'_i + \Delta) \tag{2.28}$$

$$= \int_{x'_i}^{x'_i+\Delta} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{2.29}$$

for $i \in \{-\infty, \ldots, \infty\}$ a countably bi-infinite index, and setting Δ to some suitable constant (usually based on the particular sample being considered). We recursively define $x_{i+1} = x_i + \Delta$ and $x_{i-1} = x_i - \Delta$ for some arbitrary x_0, and so we have a discretised ordering x_i over the original (continuous) domain x of $p(x)$. This discretised Gaussian distribution is interpreted as the probability that the random variable x lies in the i^{th} interval $[x_i, x_i + \Delta)$. See Fig. 2.3 for an illustration.

2.5.5 Multivariate Gaussian Distribution

Many distributions have a multivariate counterpart to their univariate version. The multivariate Gaussian is just such a counterpart to the (univariate) Gaussian described above. For the most part the multivariate Gaussian is a direct matrix gen-

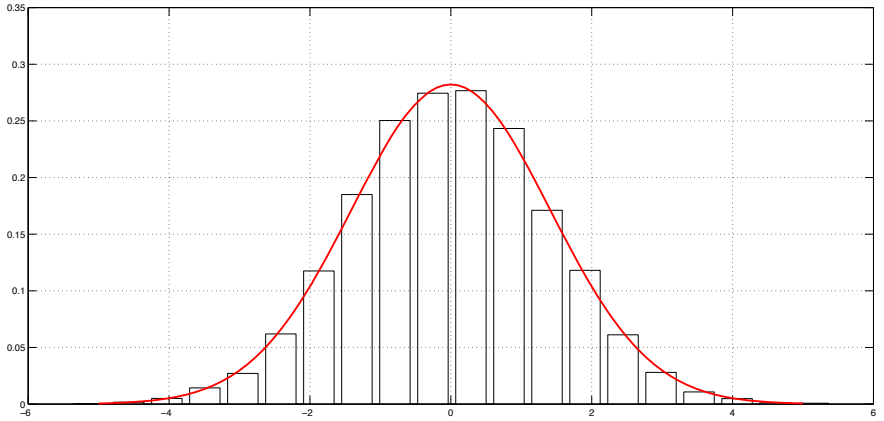


Fig. 2.3 A continuous Gaussian distribution (red) and one possible discretisation (bars)

eralisation of the univariate case, with the exception of the covariance between one component distribution with another.

The multivariate Gaussian for n random variables x_i is given by

$$p(x_1,\ldots,x_n) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)^T], \quad (2.30)$$

where Σ and $|\Sigma|$ are the covariance and the determinant of the covariance matrix, $\mathbf{x} = [x_1,\ldots,x_n]$ is the vector of random variables, $\mu = [\mu_1,\ldots,\mu_n]$ is the vector of the means of each x_i and $\mathbf{x}^T$ is the transpose of a vector $\mathbf{x}$. Explicitly, Σ must be invertible in order to form Σ^{-1} and is defined as

$$\Sigma(\mathbf{x}) = \begin{bmatrix} \mathbf{v}(x_1) & \Sigma(x_1,x_2) & \cdots & \Sigma(x_1,x_n) \\ \Sigma(x_2,x_1) & \mathbf{v}(x_2) & \cdots & \Sigma(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(x_n,x_1) & \Sigma(x_n,x_2) & \cdots & \mathbf{v}(x_n) \end{bmatrix}. \quad (2.31)$$

The special case in which $\rho(x_i,x_j) = 0 \forall x_i,x_j, i \neq j$ has a covariance matrix

$$\Sigma(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}. \quad (2.32)$$

To illustrate the multivariate Gaussian we explicitly write out the bivariate case with correlation coefficient ρ:

$$p(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_1-\mu_1}{\sigma_1^2}+\frac{x_2-\mu_2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]. \quad (2.33)$$

For $\rho = 0$ this expression reduces to $p(x_1,x_2) = p(x_1)p(x_2)$, where $p(x_1)$ and $p(x_2)$ are univariate normal distributions, as we should expect for two distributions that are (linearly) independent of one another, see Fig. 2.4. For any $\rho \in [-1,1]$ the covariance matrix for the bivariate Gaussian is:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \quad (2.34)$$

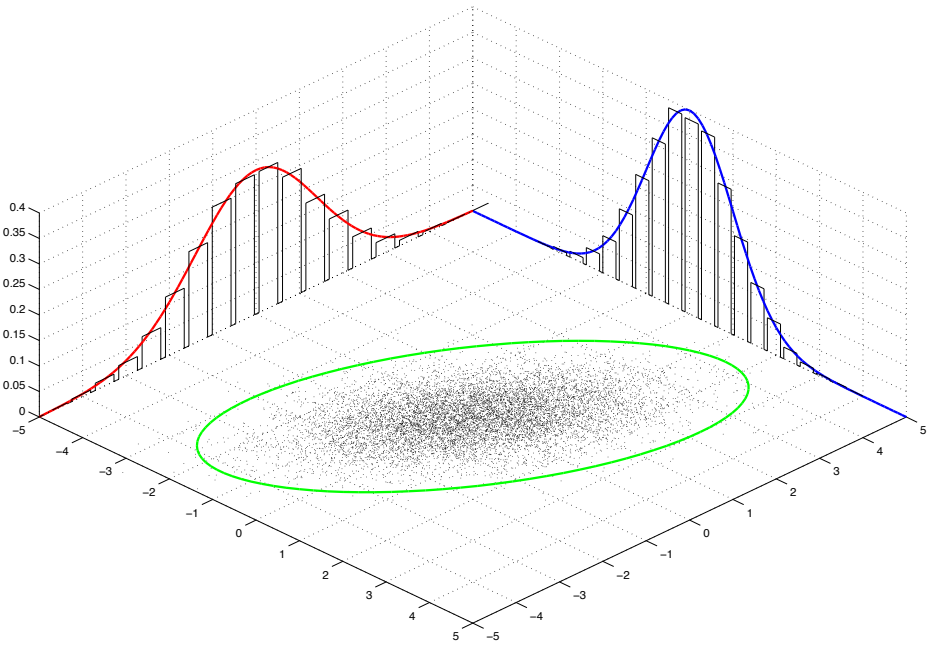


Fig. 2.4 Two coupled Gaussians with a correlation coefficient $\rho = 0.75$; the marginal probability distributions and the discretised and normalised histograms are projected onto their respective "rear walls" of the plot

2.6 Symmetry and Symmetry Breaking

The notion of symmetry and its role in theoretical physics is one of the most influential ideas to have been developed for physics and has been transferred to other sciences with exceptional success (good introductory works include [7] and [308]). Through the work of Emmy Noether, one of the most influential physicists of the 19th century and a female in a male-dominated field, a remarkable connection was established between the symmetries of a system and the conservation laws of physics such as energy and momentum.

To illustrate the notion of broken symmetry we will use the example of a potential function parameterised by the value μ as shown in Fig. 2.5. In this example a ball can be placed somewhere on the Q axis for a given value of μ, and the ball moves to either the left or the right in response to the local surface gradient of the potential given by $\phi(Q) = Q^4 + \mu Q^2$, i.e. if the gradient $\frac{d\phi}{dQ} = 4Q^3 + 2\mu Q < 0$ at $Q = 1.2$ then the ball will roll to the right, in the positive direction of Q. A *local* minimum or maximum of $\phi(Q)$ is found by solving $\frac{d\phi}{dQ} = 0$. For $\mu > 0$ there is only one point at which this occurs: $Q = 0$, but for $\mu < 0$ there are three solutions: $Q = 0$ and two other symmetrical points on either side of $Q = 0$. The $Q = 0$ solution is stable for $\mu > 0$ but unstable for $\mu < 0$. To see this, imagine a ball sitting on the $Q = 0$ point for $\mu = 4$ (i.e. at the back of the plot); small variations in the position Q of the ball, usually thought of as *thermal fluctuations*, will result in the ball returning near to the point $Q = 0$. However, for a ball placed at $\mu = -5$ and $Q = 0$, i.e. on top of the ridge at the front of the plot, a small variation in Q will result in the ball rolling down to the bottom of one of the two hollows to either the left or right of $Q = 0$, so we say that $Q = 0$ is an unstable solution to $\frac{d\phi}{dQ} = 0$. However, once the ball is in one of these two hollows, a *small* thermal fluctuation in the ball's position will result in the ball returning back to the bottom of the hollow. Both of these hollows are therefore stable solutions of $\frac{d\phi}{dQ} = 0$ for $\mu < 0$ just as the $Q = 0$ point is a stable solution of $\frac{d\phi}{dQ} = 0$ for $\mu > 0$.

There are other examples of these types of symmetries that are *spontaneously* broken as a parameter such as μ is slowly varied. A very important class of such symmetry breaking comes from the study of the Ising model used as one of the prototypical systems of purely *locally* interacting elements in a system that shows complex *global* behaviour, see Sect. 5.2 as well. The physical details are not important to the current discussion except that an important class of solutions is called the *mean-field* solution. The two-dimensional Ising model is composed of a two-dimensional grid of *particles*, each of which is connected (interacts) with four local neighbours on the grid. Each of these particles can take on one of two possible states, spin up $(+1)$ or spin down (-1), and they randomly fluctuate between these two states; the average activity of these fluctuations is called the Ising model's *temperature* T, however it is often convenient to talk about the inverse temperature $\beta = 1/T$. One of the key properties of the Ising system that we often want to understand is the average spin, or magnetisation, of the system, which is often given the variable name $Q \in [-1, 1]$; the range of values that Q can have is evident from the fact that the

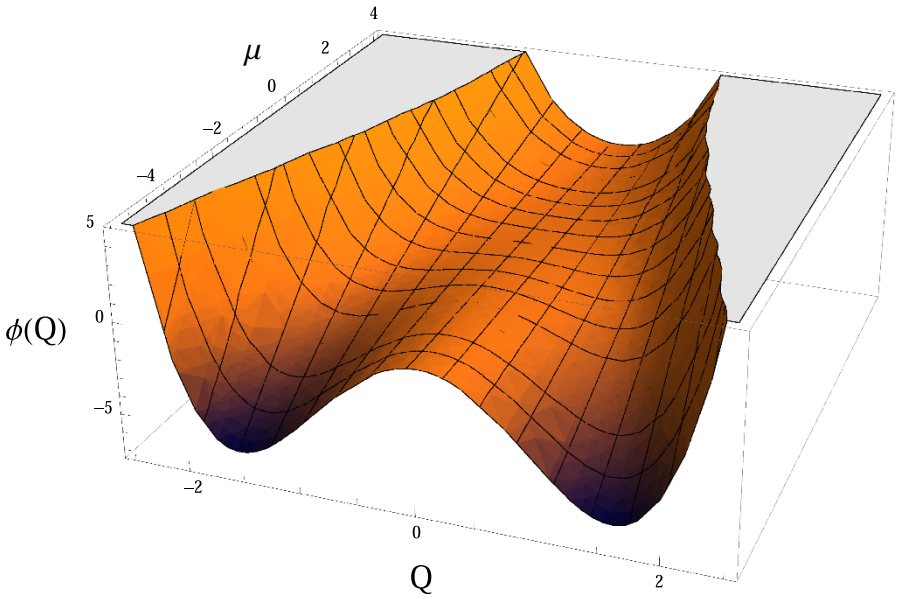


Fig. 2.5 A potential function described by $\phi(Q) = Q^4 + \mu Q^2$

average spin cannot be larger than $+1$ when all the spins are positive or smaller than -1 when all the spins are negative. It has long been known that, as the inverse temperature β varies from very small values near zero (i.e. high temperature when the spins flip very rapidly from one state to another), the average magnetisation fluctuates around $Q = 0$. However, as β increases (the system "cools"), there comes a point when the spins *spontaneously magnetise* so that $Q \neq 0$ and that as β continues to increase Q continues to diverge from $Q = 0$. The mean-field equation describing the equilibrium points of the Ising model is given by $Q = \tanh(\beta Q/2)$, solutions in terms of Q to this self-consistent equation given the mean-field approximation to the fixed points of the original Ising system, as shown in Fig. 2.6. Typically, this equation represents the equilibrium solutions to a dynamical system that evolves over time, and the long-term dynamics will be attracted to the nearest *stable* equilibrium solution and be repelled away from an *unstable* equilibrium solution. In stochastic systems, the upper and lower branches of the bifurcation shown in Fig. 2.6 might be stable while the central branch $Q = 0$ is unstable (i.e. if the system is currently at $Q = 0$ and $\beta \gg 2$ a small perturbation to Q, for example $Q = 0.01$, will result in the system evolving towards the stable branch of Q near +1).

From these examples we are able to see what is happening when a symmetry "breaks". Before the symmetry breaking point (i.e. $\beta < 2$), if we were to follow the path of a test particle as it moved around the system (pushed as it is by the thermal fluctuations), it can plausibly visit the whole system; there are no physically allowed states (or regions) of the system that are excluded to any test particle we might

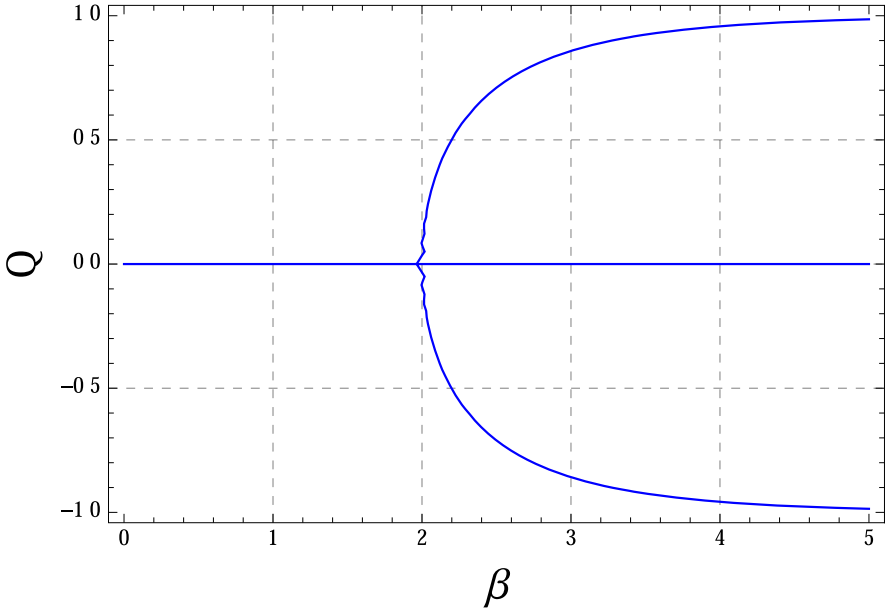


Fig. 2.6 A bifurcation plot of the equilibrium solutions to the equation $Q = \tanh(\beta Q/2)$ showing that, as β varies, the number of solutions changes from one ($\beta < 2$) to three ($\beta > 2$). The blue lines represent the expected activity (mean magnetisation) of the system; around each point of the blue lines there will be some minor thermal fluctuations

choose. However, after symmetry breaking ($\beta > 2$), there will be regions that a test particle is physically allowed to be in, but if we were to follow such a particle, we would find that some parts of the system are never explored by our test particle. We can see this by looking at Fig. 2.5 again; a test particle that starts in the left-hand hollow ($\mu < 0$) will remain in that hollow and no small fluctuations can disturb it from the left-hand hollow so that it might spontaneously jump to the right-hand hollow (it might be theoretically possible to wait for a suitably large fluctuation that *could* kick a test particle from one hollow to the other, but such fluctuations typically occur at time intervals greater than the age of the universe, essentially making the alternative hollow impossible for our test particle to visit).

So keeping in mind that the alternative hollow is still a physically allowable state but our test particle will never get there, we can introduce the notion of *ergodicity* and *broken ergodicity*, concepts that generalise the notion of symmetry and broken symmetry. An ergodic process is any statistical process that allows us to equate the time average of a test particle with the average over all allowable states of the system. For example if we had $n = 1,000,000$ observations of the Dow Jones share market index D_n, then you might think that taking this average might tell you something about the distribution of values you can expect the Dow Jones index to take (a million data points would be considered a large enough sample size for statistical

confidence). But the Dow Jones time series can be thought of as just one realisation of many possible time series (one test particle) that we can follow; how do we know that it represents a complete exploration of possible states the market can be in? How do we know that, if an underlying system parameter changes, it will not suddenly make possible other previously unseen values of the Dow Jones, values that are possibly very bad? We cannot easily know, and we often have to assume that the single time series we have been watching for a while is not stuck in a local hollow like those in Fig. 2.5.

So if the system can be stuck in one of these hollows and is unlikely to ever get out and visit the other hollow, why should we worry? The answer lies in the non-stationary nature of many complex systems. We can see how this can mislead our intuition by looking at Fig. 2.5 again and thinking in terms of two potential states of a financial market. Let us call the left well $Q-$ and think of it as the state of the market generally trending down and the right well $Q+$ and think of this as the market generally trending up; $Q-$ is called a bear market, and $Q+$ is called a bull market. An analyst watching the market increase by 9% per annum over the last few years might conclude that the market is "bullish" and the expected return on a broad portfolio of stocks will return about 9% each year (with some small variations analogous to "thermal fluctuations"), however as the economic climate changes, the mining sector, on which the economy of the country in which our analyst lives depends, begins to under-perform because the country's trading partners no longer need as much steel. This has a follow-on effect in the retail industry as jobs are lost in the mining sector and people buy less; equally the manufacturing sector struggles as there is less investment in the manufacturing of mining machines and allied industries, and the subsequent loss of employment leads to further losses in the retail sector. Ultimately the whole economy slows down, and there is no average increase in the value of the share market as many key industries are no longer growing; i.e. we reach the point where $Q = 0$ in Fig. 2.5. Now the share market begins to decrease in value every year as the circular process of economic slow down and job loss feeds back upon itself, and the same analyst now looks at the average of the share market and sees that it is now in position $Q-$, i.e. a bear market (again, enough data can be collected so that the analyst feels statistically confident that he has an accurate read on the market). In fact, if it were possible to take an average across all possible financial paths (all theoretically possible test particles), it might be that our analyst finds that the share market has an average return of a much smaller amount than he initially suspected (this is equivalent to taking averages across a statistically large number of realisations of both $Q+$ and $Q-$).

So what has happened here? We conceptualise this situation as having a macro-economic parameter μ that is driving the system dynamics and has been varying so that the single test particle the analyst was following (the time series of the market performance index that was in state $Q+$) was not representative of all of the types of states the system could be in, and so his statistical estimates were flawed despite collecting a lot of data. It is important to note that, if the macro-economic parameter μ had not varied, then the system would not have been able to switch from one broken-symmetry solution at $Q+$ to the other at $Q-$. Obviously, real economies

and real share markets are far more complex than this toy example suggests, but it does illustrate the types of complexity that we are confronted with when we (often implicitly) assume that a time series is ergodic and equate the time average of a single time-series realisation with a total exploration of the possible states of the system dynamics.

Chapter 3
Information Theory

Having cleared statistical preliminaries out of the way, we can now begin the two fundamental theoretical chapters, containing most of the mathematics. In this chapter, we introduce the fundamental concept of entropy and go on to consider mutual information on which transfer entropy is based. This chapter is somewhat more intuitive, less formal and easier to understand at a first reading than the next chapter, which gives the full mathematical details of transfer entropy.

3.1 Introduction

Entropy is one of the most alluring and powerful concepts in the history of science and information. But it initially appeared twice, largely independently. Back in the 19th century, Rudolf Clausius came up with the term in thermodynamics. Nearly a century later, Claude Shannon introduced the idea for communications and his new ideas of information theory, now fundamental to all things computational [304]. He reputedly selected this name following a suggestion from computer pioneer John von Neumann; according to Tribus [327]:

> The same function appears in statistical mechanics and, on the advice of John von Neumann, Claude Shannon called it 'entropy'. I talked with Dr. Shannon once about this, asking him why he had called his function by a name that was already in use in another field. I said that it was bound to cause some confusion between the theory of information and thermodynamics. He said that von Neumann had told him: 'No one really understands entropy. Therefore, if you know what you mean by it and you use it when you are in an argument, you will win every time.'

The thermodynamics story does not really concern us here. But the idea was simple: that systems vary in the amount of order they contain, and as they transform from one to another, this level of order is a key driver of the change. Commonly we think of entropy as the amount of disorder in a system.

Our real concern in this book is with entropy in information science, the main subject of this chapter. Entropy is the *average uncertainty* in the value of a sample

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_3

of a variable, equivalent to the *average information* required to determine the value of that sample. This begs the question of what information really is, which Sect. 3 tackles in depth. To get at information flow, a central theme of this book, we need two additional ideas: relative, or conditional entropy (CE), is what is left over when we already know something about a variable; conversely *mutual information* (MI) is essentially how much information is shared by two variables.

It is very easy to fall into a trap here, of thinking in some way about the information of individual things, say the information in a book. There is a whole field devoted to such individual information, known by the names of its three more or less simultaneous independent inventors, Kolmogorov, Solomonoff and Chaitin, and also called *algorithmic information* [179] (Sect. 3.2.6). In essence this is the length of the shortest computer program, in bits, of a description of something. The Shannon information, which is what we are interested in here, is a statistical quantity, based on sets of things. There is a link to algorithmic information theory, which we note in Sect. 3.2.6.

So, let us consider a simple example: the cars in a shopping centre car park. Given all the types of cars available, there are lots of ways of filling up the car park. This suggests that, if we know what types of cars and the numbers of each present in the supermarket car park on some given day and time, that is quite a lot of information, since it distinguishes this from all the other possible sets of cars. Now we could narrow this down a bit, if we knew something about the suburb. If it is a rich suburb we might find that there are more BMWs and Jaguars than average. If we took all the possible suburb affluence levels and all the car park contents, we would find that they were correlated, or that they have some shared, or mutual information. Sect. 3.2 makes these ideas precise and quantitative.

The driving force behind Shannon's work was coding of information to transmit it down channels, which could be noisy. Coding is not very central to this book, and there are many excellent books available. We will see occasional references to code length, but pursuing the car example will give us a bit of an intuitive insight. In the era of text messaging and twitter, we are used to abbreviating words and phrases. We are stripping a lot of the *redundancy* from English. So, in our high-class suburb, we do not need very many letters to text a new expensive car in the supermarket car park. J would do for Jaguar, L for Lamborghini (because a Lada would be very unlikely), and we would need BM for BMW, because B could also be Bentley.

But even in our high-class suburb, BMWs are more common than Lamborghinis. So we would really like to have the one-letter code for BMW and the two-letter code for the less common Lamborghini. With text messages, and social media such as Twitter, we are now quite happy with using short abbreviations for terms and phrases which occur frequently. Coding theory provides ways of doing this in the most economical way, the so-called optimal codings.

Shannon published his ideas on information in the 1940s. For quite a long time after that, all practical uses relied on simplifications, such as the assumption of Gaussian processes. The reason for this is partly that these quantities are hard and expensive to calculate, i.e. they take a lot of computing time. Sect. 3.4 takes a look at why this is so, and surveys some of the algorithms we can use to get better es-

timates. It is interesting that some questions, such as the mutual information of the Ising model, discussed in Chap. 5, are still seeing solutions published in 2013, decades after the models first appeared. There are many good books on information theory, such as MacKay [208], which also covers coding and Bayesian statistics.

3.2 Basic Ideas

3.2.1 Entropy and Information

Entropy as a thermodynamic concept is a measure of the disorder in a system. In fact, the phase transitions we shall talk about in this book (Sect. 3.3) are transitions between different levels of order, between order and disorder, or vice versa. In information theory there is an additional way of looking at entropy: it is the average uncertainty in the value of a sample of a variable, equivalent to the average information required to predict the value of that sample. To understand this we need to go back to the definition of information itself.

Shannon wanted an information measure which satisfied a number of conditions, notably:

1. It should be additive for independent pieces of information
2. It should reflect likelihood of events, in particular capturing increasing uncertainty associated with an increasing number of (equally likely) events.
3. It should be continuous with respect to changes in these likelihoods.

He was interested in how much information a message conveyed. If something is very likely to happen, the information gleaned from it happening is not very great, a bit like the sun rising in the morning does not actually tell you very much. On the other hand, rare events (such as the sun shining while it is raining) convey a great deal of information because they are relatively surprising. Thus, his measure of the information, $\eta(x)$, of an event x, was the log of the probability, $p(x)$, of x happening, being observed, whatever (Eqn. 3.1). This is also sometimes called the *surprisal* or *Shannon information content*. Formally, following the nomenclature of Chap. 2, we consider samples of a random variable x of the event X, which take values from a discrete alphabet or space of all elementary events Ω_X as described by a probability function $p(x)$ (Sect. 2.1) with total number of events $M = |\Omega_X|$. We omit the subscript X where no ambiguity results. In cases where it is not explicitly specified, a summation over x implies one over Ω_X.

$$\eta(x) = -\log_2 p(x). \tag{3.1}$$

Shannon used natural logs, giving information in *nats*. When we consider Gaussian variables, natural logs appear directly, but in most cases we shall use logs to base 2, denoted $\log_2$, giving information in bits, the more common unit today.

One can interpret the values of $\eta(x)$, in bits, as the optimal number of yes/no questions that one needs to ask (on average) to determine the value of x. For example, say I want to know the make x of one particular car in our supermarket parking lot, and you already know the answer. If I ask "Is it a UK make?" and you say "Yes", that narrows it down to a Jaguar or Bentley. I can then get the final answer by asking "Does it begin with "J?", and let us assume you say "Yes". Assuming that we had a 50-50 probability for the answer for each question, then the Shannon information content for $x =$ Jaguar here was 2 bits, corresponding to two yes/no questions. Another interpretation in terms of optimal code lengths follows later in Sect. 3.2.7

Because information is an exceptionally common idea and a term in very common usage in today's world, we have to be very careful not to confuse everyday parlance with the mathematically precise concept we need for this book. It is very easy to get mixed up with the vernacular ideas of information in something (like the newspaper), the content itself of information sources, the data on a hard drive and so on.

Key Idea 2: *The information of information theory has nothing to do with meaning.*

It is very important to be clear about context. Information can mean lots of things. A lion in the garden in England suggests a zoo or wildlife park has some escapees. A lion in the garden in Kenya may not mean very much at all, other than to stay indoors. We have to strip away all extraneous factors, and consider sets of events. Information generalises to continuous probabilities, but there are some mathematical niceties (Sect. 3.2.5 and Fig. 3.1).

Key Idea 3: *Shannon information is a property of sets of objects, not the objects themselves.*

Given this definition of information, the entropy is now the *average information* over sets of events, which can be measured as repeated observations over time, or over sets of different realisations of a system, the two being equivalent when the system is ergodic (Sect. 2.6).

If we average or take the expectation value (Sect. 2.4) of the information according to the probability of each event occurring, we end up with the Shannon *entropy*, Eqn. 3.2.

$$\mathbf{H}(X) = E\{\eta(x)\} = -\sum_{x\in\Omega} p(x)\log_2 p(x). \tag{3.2}$$

This is a quite general principle for all the entropy and information measures in this book. We can get a system-level descriptor, such as the entropy, by averaging over all the descriptors for just a single point or event. These single-point descriptors, we call *pointwise* descriptors, but they are sometimes called *local*, i.e. *local*

entropy.[1] This suggests using a unified notation: lower case is used for the pointwise descriptor and upper case for the system level. Thus, we rewrite $\eta(x)$ as $\mathbf{h}(x)$.

Key Idea 4: *All the system-level information-theoretic quantities may be expressed as expectation values over the pointwise (local) quantities.*

To add subtlety to Key Idea 3, we clarify that Shannon information content, or local entropy, is a property associated with each object—but only in the context of the whole set of objects.

For two variables, the *joint entropy* is simply

$$\mathbf{H}(X,Y) = -\sum_{x\in\Omega_x}\sum_{y\in\Omega_y} p(x,y)\log_2 p(x,y), \tag{3.3}$$

and so on for any number of variables.

We also need the idea of *conditional entropy*, the uncertainty left after we have taken into consideration some context. So, if we look at the different brands of cars we would see on a typical high street, then new Ferraris would be quite rare. But if the high street is in Hampstead in London, or the centre of Dubai, there may be lots of very expensive cars. So, if X is the set of car brands and Y the set of high streets, we determine the entropy of X for each high street y, and it will vary widely with location, Eqn. 3.4

$$\mathbf{H}(X|y) = -\sum_{x\in\Omega_x} p(x|y)\log_2 p(x|y). \tag{3.4}$$

To get the conditional entropy, we just have to average over the different high streets, Eqn. 3.5:

$$\mathbf{H}(X|Y) = \sum_{y\in\Omega_y} p(y)H(X|y). \tag{3.5}$$

We can write a *conditional Shannon information content* (or local conditional entropy) for the information content of event x given that event y occurs:

$$\mathbf{h}(x\,|\,y) = -\log_2 p(x|y), \tag{3.6}$$

$$\mathbf{H}(X\,|\,Y) = E\{\mathbf{h}(x\,|\,y)\}\,. \tag{3.7}$$

We can then rewrite Eqn. 3.3 in terms of the conditional entropy (Eqn. 3.8). This simply states that we take the entropy of X, then add what is left of the entropy of Y after taking out any X dependence, or vice versa:

[1] A word of caution is needed here. Local entropy and local mutual information are sometimes used elsewhere to mean computation of entropy, say, over a local area of samples, rather than pointwise at a given sample.

$$\mathbf{H}(X,Y) = \mathbf{H}(X) + \mathbf{H}(Y|X) = \mathbf{H}(Y) + \mathbf{H}(X|Y). \tag{3.8}$$

Fig. 3.1 Low- and high-entropy fur! How would you interpret the entropy of feline fur? There is no one answer to this. Is it meaningful to talk about the fur entropy of the calico cat? Where does the giant statue in Barcelona fit in?

3.2.2 Mutual Information

The *mutual information* is the amount of *shared* information between X and Y. It is a measure of their *statistical dependence* (Sect. 2.3.2). Thus, we should be able to take the entropy of X and subtract from it the entropy of X given Y, since this chunk of the entropy has, by definition, nothing to do with Y. This is exactly the case as in Eqn. 3.9.

The mutual information can be thought of as a non-linear form of correlation (Sect. 2.4).[2] The corollary of this is that

$\mathbf{I}(X:Y) = 0 \iff X$ is independent of Y.

$$\mathbf{I}(X:Y) = \mathbf{H}(X) - \mathbf{H}(X|Y) = \mathbf{H}(Y) - \mathbf{H}(Y|X), \tag{3.9}$$

which is clearly symmetric in X and Y.

Instead of thinking about subtracting out the conditional component, we can start with the marginal entropies, $\mathbf{H}(X), \mathbf{H}(Y)$. If there is any shared information, the sum of these should be bigger than the joint entropy, by the MI:

$$\mathbf{I}(X:Y) = \mathbf{H}(X) + \mathbf{H}(Y) - \mathbf{H}(X,Y). \tag{3.10}$$

The MI must therefore always be non-negative. In terms of probabilities, we have

$$\mathbf{I}(X:Y) = \sum_{x\in\Omega_x, y\in\Omega_y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \tag{3.11}$$

This expression is a particular example of the Kullback–Leibler divergence (KLD, see Sect. 3.2.4), a measure of the information gap between dependence and independence.

Key Idea 5: *Mutual information is the total marginal entropy minus the joint entropy, or the Kullback–Leibler divergence of the product of marginal distributions from the joint distribution.*

We can also use the pointwise (local) mutual information between specific events x and y [88]:[3]

$$\mathbf{i}(x:y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{p(x \mid y)}{p(x)}, \tag{3.12}$$

$$\mathbf{I}(X:Y) = E\{\mathbf{i}(x:y)\}. \tag{3.13}$$

[2] For non-binary, non-Gaussian variables, the mutual information can be large while the correlation is low, and vice versa. For binary variables, zero correlation does imply independence [267].

[3] Fano [88] demonstrated the uniqueness of this form under a set of axioms from which the mutual information is derived. This is in contrast to *partially* localised mutual information expressions, $\mathbf{I}(X:y)$ (also known as *specific information*), which consider how much information a specific event y provides on average about the other unknown variable X, of which there are two possible forms satisfying different criteria [78].

3.2.2.1 Misinformation

Importantly, the pointwise mutual information may be either positive *or negative* for a specific pair x, y. Positive values are easy to understand, occurring where $p(x \mid y) > p(x)$, i.e. knowing event y *increases* our expectation of the occurrence of event x. Negative values simply occur in Eqn. 3.12 where $p(x \mid y) < p(x)$. That is, knowing event y changed our belief $p(x)$ about the probability of occurrence of event x to a smaller value $p(x \mid y)$, and hence we considered it less likely that x would occur when knowing y than when not knowing y, in a case where x nevertheless occurred. We can say that y was *misinformative* about the value of x. So, imagine we have two ethnic quarters in a city. In one, A, 90% of people have blond hair and 10% black. In the other, B, 90% have black hair, the rest blond. So meeting somebody from outside one of these suburbs, there is a 50% chance of either colour. But if we are told that somebody is from A, then we would expect them to be blond, so the pointwise MI that they have black hair is negative. We give a more detailed example in Sect. 3.2.2.3 below.

3.2.2.2 Multi-information: Mutual Information for Three or More Variables

But why stop our consideration of shared information at two variables? Surely there could be some common information amongst any number of variables. The generalisation of MI to more than two variables from Eqn. 3.10 is easy, giving the *multi-information* or *integration* in Eqn. 3.14 [326]. There are numerous practical situations where we might want to calculate the mutual information amongst numerous variables. So in Chap. 6 the peak in mutual information amongst stock prices in an index during a crash is considered. Stock market indices may have many stocks, from the Dow Jones which has 30 to the S&P 500, which, as you might expect, has 500. But the multi-information can also be used to describe transfer entropy, as we shall see in Chap. 4, although using MI may not be a good way of estimating TE from data. The definition of multi-information is not immediately obvious though, since mutual information was originally defined to be between just two things, and thus, there is more than one proposal; MacKay [208] considers the three-term MI of Eqn. 3.14 to be illegal. The definition of Eqn. 3.14 is a straightforward generalisation of the Kullback–Leibler form, and was used, for example, by Fraser and Swinney in studying entropy of chaotic attractors [92].

$$\mathbf{I}(X_1 : X_2 : X_3 : ...X_n) = \mathbf{H}(X_1) + \mathbf{H}(X_2) + \mathbf{H}(X_3)... + \mathbf{H}(X_n) - \mathbf{H}(X_1, X_2, X_3, ...X_n). \tag{3.14}$$

We can also expand the multi-information using similar thinking. The three-way MI ought to be MI of any given pair, plus the MI of the third variable with this pair. So it turns out to be

$$\mathbf{I}(X_1 : X_2 : X_3) = \mathbf{I}(X_1 : X_2, X_3) + \mathbf{I}(X_2 : X_3). \tag{3.15}$$

As the number of variables increases, the computational load and data demands increase dramatically. Thus, using pairwise approximations is highly desirable. Unfortunately this does not always work as well as one would hope [288], and each case needs to be carefully assessed.

3.2.2.3 A Health Diet

The media constantly regale us with directives to eat more fruit and vegetables. But the food and botanical worlds differ in what they call a fruit and what they call a vegetable. Avocado and tomato, for example, are technically fruits. Such classifications come about by sharing properties. In the case of food, the dominant property is salty/sweet. In the botanical world, the presence of a stone or seeds defines a fruit. So there are overlapping properties, and thus some mutual information among them. Thus, we are going to calculate the multi-information for culinary type (fruit, veg), colour (yellow, green, black) and stoned (yes or no), see Table 3.1. Our set of fruit and veg (events for occurrence of these properties), all equally likely, is: yellow fruit – grapefruit, lemon, banana, apricot, peach; green fruit – melon, gooseberry; black fruit – cherry, damson; black veg – aubergine; green veg – avocado, kale, rocket, courgette; yellow veg – turnip, onion. There are 17 items. We can calculate occurrences and hence probabilities for each in Table 3.1.

Type Stoned	Fruit Yes	Fruit No	Veg Yes	Veg No
Yellow	2(A,P)	3(Gf,L,B)	0	3(T,S,O)
Green	0	2 (M,G)	1(Av)	3(K,R,Z)
Black	2(D,C)	0	0	1(E)

Table 3.1 Fruit and vegetable occurrence table

From this table we calculate the entropy of type (0.998), colour (1.484) and stoned (0.874) and the joint entropy (2.895) and thus, the multi-MI (0.460).

We also have the mutual information between *types (T): fruit (f)* and *vegetable (v)* and *Stone (S): stoned (s)* or *not-stoned (n)*: $\mathbf{I}(T:S) = 0.093$ bits. Looking at local MI values (Eqn. 3.12), we have $\mathbf{i}(f:s) = 0.596$ bits, which is positive since $p(f \mid s) = 0.8 > p(f) = 0.529$, whereas $\mathbf{i}(v:s) = -1.235$ bits is negative since $p(v \mid s) = 0.2 < p(v) = 0.471$. This means that, if we have an avocado (a stoned vegetable), then knowing it is stoned actually *misinforms* us about its status as a vegetable, since we would expect it to be a fruit.

3.2.3 Conditional Mutual Information

An important generalisation of mutual information, which is crucial to the development of transfer entropy (Chap. 4) is the idea of mutual information between two processes, X and Y, *conditioned* on a third process, Z.

Imagine that we start to see snakes and an increasing number of cats in the garden. Now cats do kill snakes, and we might think that the cats are there just for that purpose. However, killing a snake is just a bit risky for a cat. But we might find that the snakes and cats appear because of a third factor—an increase in the number of mice. If we condition on the mice population, then we find there is no relationship between cats and snakes in the garden; they are *conditionally independent* (Sect. 2.3.4).

In fact the expression for the *conditional mutual information*, $\mathbf{I}(X:Y|Z)$, is very straightforward. We simply condition each of the entropy terms in Eqn. 3.9:

$$\mathbf{I}(X:Y\,|\,Z) = \mathbf{H}(X\,|\,Z) - \mathbf{H}(X\,|\,Y,Z)\,. \tag{3.16}$$

or

$$\mathbf{I}(X:Y\,|\,Z) = \mathbf{H}(X\,|\,Z) + \mathbf{H}(Y\,|\,Z) - \mathbf{H}(X,Y\,|\,Z)\,. \tag{3.17}$$

with the following conditional independence criterion

$\mathbf{I}(X:Y\,|\,Z) = 0 \iff X$, conditional on Z, is independent of Y.

Returning to our cats-snakes-mice example, we may measure some mutual information between the population of cats X and population of snakes Y. But if we condition on the mice population Z, then we find that the conditional MI is zero, since the population of cats *conditional on the mice population* is independent of the population of snakes.

Again we can write the conditional mutual information as an average or expectation value over the pointwise quantity, Eqn. 3.19 [88]:

$$\mathbf{i}(x:y\,|\,z) = \log_2 \frac{p(x\,|\,y,z)}{p(x\,|\,z)}, \tag{3.18}$$

$$\mathbf{I}(X:Y\,|\,Z) = E\{\mathbf{i}(x:y\,|\,z)\}\,. \tag{3.19}$$

While the conditional MI $\mathbf{I}(X:Y\,|\,Z) \geq 0$, the pointwise conditional mutual information $\mathbf{i}(x:y\,|\,z)$ may be either positive *or negative* for a specific event $\{x,y,z\}$, as per the local (unconditioned) mutual information.

3.2.3.1 Redundancy and Synergy

At first it seems as if conditioning "out" some other variable should make the mutual information decrease (as occurs for entropies). But this is not always the case. A conditional MI $\mathbf{I}(X:Y\,|\,Z)$ may be either *larger or smaller* than the related unconditioned MI $\mathbf{I}(X:Y)$ [208]. Such conditioning removes *redundant* information in Y

and Z about X, but also adds *synergistic* information which can only be decoded with knowledge of both Y and Z.

Example 3.1. Redundancy: Where we have $X = Y = Z$ for random either/or events, such as coin flips, then $\mathbf{I}(X:Y) = \mathbf{I}(X:Z) = 1$ bit. However $\mathbf{I}(X:Y \mid Z) = \mathbf{I}(X:Z \mid Y) = 0$ because Y and Z redundantly hold the same information about X. If we know Y then we do not get any more information by learning about Z.

Example 3.2. Synergy: The classic example of synergy is a Boolean exclusive OR (XOR) operation $X = Y$ XOR Z (see Table 3.2). When Y and Z are independent and randomised, then $\mathbf{I}(X:Y) = \mathbf{I}(X:Z) = 0$, however conditioning on the other input reveals the synergistic relationship and we have $\mathbf{I}(X:Y \mid Z) = \mathbf{I}(X:Z \mid Y) = 1$ bit.

Table 3.2 Exclusive OR (XOR) Boolean operation $X = Y$ XOR Z. Resulting values for X are listed in the logic table for each Y, Z pair

		Y	
		0	1
Z	0	**0**	**1**
	1	**1**	**0**

Crucially, these redundant and synergistic components can occur simultaneously (unlike in the examples above—see descriptions of OR and AND logic gates in [121]), and they cannot be measured with classic information-theoretic terms. Several significant efforts are ongoing to attempt to measure these quantities, broadly termed the *partial information decomposition* approach [358, 359, 131, 121, 186, 35, 36, 325].

Open Research Question 1: *How should synergy and redundancy components of mutual information from a set of sources to a target be properly measured? Indeed, is this possible in general, or only in limited circumstances?*

3.2.4 Kullback–Leibler Divergence

The *Kullback–Leibler divergence* (KLD) measures the information required to tell one probability distribution, $q(x)$ from another $p(x)$:

$$\mathscr{K}(p||q) = \sum_{x \in \Omega_x} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) dx. \tag{3.20}$$

Kullback and Leibler formulated this in terms of the information for discriminating between hypotheses: for any value x, the *information in x for discrimination of the hypothesis that x is drawn from p or q*, ξ, is

$$\xi = \log_2\left(\frac{p(x)}{q(x)}\right). \tag{3.21}$$

Averaging over p now gives Eqn. 3.20.

The *cross entropy*, $\mathbf{G}(p:q)$ is defined by

$$\mathbf{G}(p:q) = -\sum_{x\in\Omega_x} p(x)\log_2 q(x), \tag{3.22}$$

and thus, the KLD can be expressed as

$$\mathscr{K}(p||q) = \mathbf{G}(p:q) - \mathbf{H}(p). \tag{3.23}$$

Alternatively, the KLD of $q(x)$ from $p(x)$, $\mathscr{K}(p||q)$, has an information-theoretic interpretation. It is the amount of information lost when using $q(x)$ to represent $p(x)$.

Thus, the MI is the KLD of the product of the marginals $p(x)p(y)$ from the joint distribution $p(x,y)$, as we have just discussed. That is, if we replace p with the joint distribution $p(x,y)$ and q with the product of the marginals $p(x)p(y)$, we end up with the continuous form of Eqn. 3.11. Note that the KLD is *not* symmetric, whereas the MI, of course, is. Just as the mutual information is always non-negative, so, the KLD is positive or zero. This follows from the *Gibbs inequality*, which states that the entropy is always less than the cross entropy with any other function:

$$H(p) \leq G(p,q), \quad \forall q, \tag{3.24}$$

with equality if $p \equiv q$.

3.2.4.1 Hot and Sticky – KLD example

Cairns, in Far North Queensland in Australia, has a *lot* of rain in the summer, when, even though Cairns is between the Equator and the Tropic of Capricorn, it is still hotter in summer than in winter. Table 3.3 shows illustrative average monthly temperature and rainfall.

So, we can ask how much information we get about the variation of temperature from knowing the rainfall. A simple way to do this is to examine the two probability distributions for binary variables, $p(\texttt{wet})$ and $p(\texttt{hot})$. Applying the formula for KLD (Eqn. 3.20) we get 0.63 bits for temperature to rain. In other words we lose 0.63 bits of information if we use rain to approximate temperature. We lose 0.53 bits for using temperature as an approximation for rain. As we can see, they are not equal. We go on to examine the MI between temperature and rainfall in Sect. 3.2.5.4.

Item	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Mean max temp.(°C)	31.4	31.2	30.6	29.2	27.6	26.0	25.7	26.6	28.1	29.5	30.6	31.4
Mean rainfall (mm)	395.3	450.6	424.2	195.1	91.4	45.3	29.5	27.0	33.7	46.6	93.8	178.8
$p(\texttt{wet})$	0.200	0.224	0.211	0.097	0.045	0.023	0.015	0.013	0.017	0.023	0.047	0.089
$p(\texttt{hot})$	0.090	0.090	0.088	0.084	0.079	0.075	0.074	0.077	0.081	0.085	0.088	0.090

Table 3.3 Cairns climate data: mean daily maximum temperature and mean monthly rainfall retrieved from the Australian Bureau of Meteorology (http://www.bom.gov.au, 18 August 2013). Mean daily maximum temperature and monthly rainfall are 29.0°C and 168 mm. $p(\texttt{wet})$ and $p(\texttt{hot})$ are illustrative constructions for the rainfall or temperature being above some threshold each day

3.2.5 Entropy of Continuous Processes

The discrete forms of entropy we have discussed so far do not go across smoothly to the case of continuous probability density functions (PDFs) (see Sect. 2.5.3), with the sum in Eqn. 3.2 going over to an integral. In fact we have to *define* the continuous entropy as the integral, Eqn. 3.25 [256], also referred to as the *differential entropy*:

$$\mathbf{H}_{cont}(X) = -\int_{-\infty}^{\infty} p(x)\log p(x)dx, \tag{3.25}$$

where the integral *excludes* points where $p(x) = 0$. Shannon also introduced this expression [304]. Of course, this may be extended to a multivariate or vector $\mathbf{x}$, with a multiple integral over each dimension. Notice that the differential entropy normally uses natural logs, and when this is the case returns units of *nats*.

Key Idea 6: *The properties of the differential entropy can be counter-intuitive in comparison with those of the Shannon entropy (of discrete variables); e.g. it can be negative.*

Shannon noted that the differential entropy is dependent on the coordinates, and may change under a coordinate transformation. For a linear transformation of the components x_i of the multivariate X we have: $u_j = \sum_i a_{ij}x_i$ for the multivariates U and X and transformation matrix A, and the change in entropy reduces to the log of the determinant of A:

$$\mathbf{H}_{cont}(U) = \mathbf{H}_{cont}(X) + \log|A|. \tag{3.26}$$

Crucially, this illustrates that differential entropies may be negative, unlike entropy of discrete variables.[4] We also note that differential entropy does not change with shifting of the coordinates, i.e.:

[4] This can easily be verified: for a univariate X with finite entropy $\mathbf{H}_{cont}(aX)$, we can always select a scaling factor a small enough to make $\log a$ a large enough negative number such that $\mathbf{H}_{cont}(aX)$ becomes negative.

$$\mathbf{H}_{cont}(X+a) = \mathbf{H}_{cont}(X)\,. \tag{3.27}$$

Key Idea 7: *Other information-theoretic terms (e.g. conditional entropies, MI and conditional MI) applied to multivariate distributions may be formed as the sums and differences of the underlying entropy terms (with each evaluated as per Eqn. 3.25).*

Key Idea 8: *Crucially, the differential MI (and conditional MI) has certain properties matching those for discrete variables (i.e. being non-negative), and does not change with scaling of the variables.*

Indeed, the differential MI is equal to an MI calculated on the discretisation of continuous variables (see Sect. 2.5.3) in the limit as the bin size approaches zero.

In a similar fashion, we can write the Kullback–Leibler divergence for continuous probability density functions:[5]

$$\mathscr{K}(p||q) = \int p(x)\log\frac{p(x)}{q(x)}dx. \tag{3.28}$$

3.2.5.1 Entropy of Gaussian Processes

Substituting the Gaussian function

$$G(x) = \frac{1}{\sqrt{(2\pi)}\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}} \tag{3.29}$$

into Eqn. 3.25 leads directly to the differential entropy, Eqn. 3.30, as a function of the standard deviation σ (or variance σ^2):

$$\mathbf{H}(\sigma) = \log\sqrt{2\pi e}\sigma, \tag{3.30}$$

in *nats* by convention for Gaussian variables.

For the k-dimensional Gaussian $\mathbf{x}$ with mean μ and covariance matrix Σ

$$G(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left((\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{3.31}$$

the entropy becomes (in *nats*)

[5] Again with the domain of the integral *excluding* points where $p(x) = 0$.

$$\mathbf{H}_{mvg} = \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|, \tag{3.32}$$

where $|\Sigma|$ is the determinant of Σ. Notice that $\mathbf{H}_{mvg}$ is independent of the mean μ, verifying the shift invariance of Eqn. 3.27.

Other information-theoretic terms (e.g. mutual information) applied to multivariate Gaussian distributions may be formed as the sum and difference of the underlying entropy terms (as per Key Idea 7, with each evaluated as per Eqn. 3.32).

Of course, we can use Eqn. 3.30 to construct simple and fast estimators of differential entropy and related quantities using the empirically determined covariances of the processes, acknowledging that this assumes an underlying (multivariate) Gaussian model of the distributions, and linear relationships between the variables.

There are two interesting limiting cases of entropy and mutual information using Gaussian distributions. If the mean and variance of a distribution are given, the Shannon entropy is maximal if the distribution is Gaussian. The next case concerns the MI, as discussed below.

3.2.5.2 Mutual Information of Gaussian Processes

For a given covariance matrix (with the important assumption of Gaussian marginals), a multivariate Gaussian distribution gives the lower bound on mutual information [91].[6]

The multi-information of the multivariate Gaussian $\{X_1, X_2 \ldots\}$ is given by

$$\mathbf{I}(X_1, X_2, \ldots X_n) \geq -\frac{1}{2}\log\frac{|\Sigma|}{\sigma_1^2\sigma_2^2, \ldots \sigma_n^2} \tag{3.33}$$

in *nats*, where Σ is the covariance matrix and σ_i^2 are the individual variances.

For the standard MI of two variables, this simplifies to

$$\mathbf{I}(X:Y) = -\frac{1}{2}\log(1-\rho^2). \tag{3.34}$$

(in *nats*), where ρ is the correlation coefficient between the variables. This may be derived using Eqn. 3.32 for each required entropy in Eqn. 3.10, and we also show an alternative derivation using KLD in Sect. 3.2.5.3.

Key Idea 9: *The MI between two Gaussian variables is completely determined by their correlation coefficient ρ in Eqn. 3.34, increasing with the magnitude of ρ.*

[6] Kraskov et al. [168] originally claimed this was the case for any marginal distribution, however this was corrected in [91].

Interestingly, we see here that, if $\rho = 1$ or -1 for completely correlated or anti-correlated variables, respectively, then the MI between them diverges. We can interpret this result in that there is infinite precision contained in the complete specification of one random Gaussian variable (since it is a real number), and if one such variable completely specifies another (with $\rho = 1$ or -1), then it must be providing an infinite amount of information about that other variable.

3.2.5.3 Kullback–Leibler Divergence for Gaussians

We can also write down an analytic solution for the differential entropy-based KLD (Eqn. 3.28) on Gaussian multivariates. For two Gaussians, x, y with mean μ_x, μ_y and standard deviation σ_p, σ_q, the KLD is given by Eqn. 3.35, using the definition of Gaussian (Sect. 2.5.4) repeated in Eqn. 3.29 (Fig. 3.2).

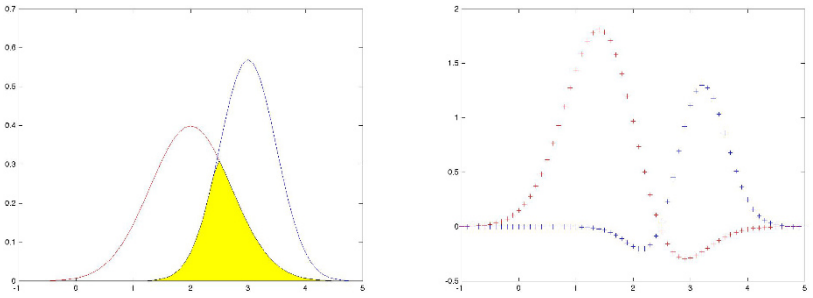


Fig. 3.2 Two Gaussian distributions with different mean and variance. The KLD depends on how much the distributions overlap, shown here as a yellow area in the left-hand figure. As the yellow area increases, as the two curves move closer, the KLD decreases, reaching zero when the curves overlap completely. To see the asymmetry in the KLD, the right-hand figure shows the integrand of Eqn. 3.20: the red curve (plus signs) is $\mathscr{K}(a||b)$ and the blue curve (circles) is $\mathscr{K}(b||a)$, where a is the curve with the maximum to the left of b

$$\mathscr{K}(x||y) = \frac{1}{2\log 2}\left\{\frac{\sigma_x^2}{\sigma_y^2} + \frac{(\mu_y - \mu_x)^2}{\sigma_y^2} - 1 - \log\left(\frac{\sigma_x^2}{\sigma_y^2}\right)\right\}. \tag{3.35}$$

If the means and standard deviations are the same, the KLD is zero as we would expect. For a multivariate Gaussian (see Eqn. 3.31), with joint covariance matrix Σ and covariance matrices Σ_x and Σ_y for each marginal variable, the KLD is given by

$$\mathscr{K}(\mathbf{x}||\mathbf{y}) = \frac{1}{2\log 2}\left\{\mathrm{tr}\left(\Sigma_y^{-1}\Sigma_x\right) + (\mu_x - \mu_y)^T\Sigma_y^{-1}(\mu_x - \mu_y) - K - \log\left(\frac{|\Sigma_x|}{|\Sigma_y|}\right)\right\}, \tag{3.36}$$

where K is the dimensionality of $\mathbf{x}$ and $\mathbf{y}$, assumed the same.

We can now use the KLD $\mathscr{K}(p(x,y)||p(x)p(y))$ to derive the MI for two correlated Gaussian variables x, y with correlation coefficient ρ, to give Eqn. 3.34, dependent only on the correlation. The covariance matrix for the joint distribution $p(x,y)$ is given by

$$\Sigma_{p(x,y)} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{3.37}$$

The covariance matrix for the product of marginals $p(x)p(y)$ is simply a diagonal matrix in the variances:

$$\Sigma_{p(x)p(y)} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}. \tag{3.38}$$

We also need the inverse of this matrix, which can be found by standard linear algebra:

$$\Sigma^{-1}_{p(x)p(y)} = \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix}. \tag{3.39}$$

To make life easy, we assume zero mean,[7] plug into Eqn. 3.36 and, after a little algebra, we get

$$\mathbf{I}(X:Y) = \mathscr{K}(p(x,y)||p(x)p(y)) = -0.5\log(1-\rho^2). \tag{3.40}$$

3.2.5.4 Hot and Sticky – MI Example

To demonstrate the MI on continuous-valued variables, we return to the climate data for Cairns in Table 3.3. Cairns is within the Tropics and has closer to a wet (summer) and dry (winter) season than the four seasons of more temperate regions. It never gets really cold, but summer is definitely warmer. So we can ask what information the temperature would tell us about whether it was summer or winter. Now the distributions of temperature for summer and winter are not necessarily the same shape, but we can see why this is. There are plenty of very hot days in summer, but none in winter. But there are cool days in summer. So temperature tells us more about summer than about winter.

Now, let us add some numbers to this idea by asking what the MI is between monthly rainfall (`rain`) and mean maximum daily temperature (`temp`) using the data in Table 3.3. Note that, whereas for the KLD estimate in Sect. 3.2.4.1 we used the probability distribution for the *discrete* variables `wet` and `hot`, we are now directly examining the numerical relationship between the *continuous* variables `rain` and `temp`. We can calculate an estimate for this MI (under a Gaussian model for the data distribution) using Eqn. 3.34: we have $\rho = 0.7315$, and therefore we estimate $\mathbf{I}(\texttt{rain}:\texttt{temp}) = 0.55$ bits. The interested reader can experiment with the other estimators presented in Sect. 3.4.2 and compare with this Gaussian model estimate.

[7] Which does not change the final value, since differential entropies are constant under co-ordinate shift – see Eqn. 3.27.

3.2.6 Entropy and Kolmogorov Complexity

One element of confusion experienced by newcomers to the ideas of entropy and information is that they have nothing to say about the complexity or information in any object. They are properties of *groups of objects,* and the only thing that matters about the object is some frequency of occurrence. They could be potatoes or pandas. But for the curious reader, there *is* a relationship.

With some caveats, Eqn. 3.2 is the average length of the shortest description of an observation of the variable x [68]. We can make this idea of description length more precise. The Kolmogorov complexity does measure the information in an object in a certain sense. It is the length of the shortest program required to describe the object. There are obviously a lot of details needed to make this idea precise, which would not only take us off course, but into very choppy waters, but the bottom line is

$$\mathbf{H}(X) = \sum_{x \in \Omega} p(x)\mathbf{K}(x), \tag{3.41}$$

where $\mathbf{K}(x)$ is the Kolmogorov complexity of object x with probability $p(x)$ from the set X. Full details can be found in the book by Li and Vitanyí [179].

3.2.7 Historical Note: Mutual Information and Communication

Shannon introduced MI for his work on the information which could be carried by a channel, noisy or not. Communication is not a central theme of this book, but Shannon's insight was iconoclastic and is worth describing briefly. One might think that, if a channel is noisy, data will always get lost sometimes. But this is not so. Shannon showed that, if the entropy of the source data being fed into the channel is less than the *channel capacity*, then there will always be a way of coding the data such that it can be perfectly decoded at the other end of the channel. We get the channel capacity by calculating the mutual information between the signal S and the noise N. If both are Gaussian, then it is easy to show that this will turn out to be

$$\mathbf{I}(S+N:N) = \frac{1}{2}(1+\frac{\sigma_S^2}{\sigma_N^2}). \tag{3.42}$$

In this book we shall not pursue the labyrinthine details of optimal coding. But it does provide another way of looking at the pointwise and system measures. Thus the entropy, $\mathbf{H}(X)$, explicitly captures the average code length (a number of bits) to encode each event x in an *optimal encoding scheme* for the measurements X, while the information, $\mathbf{h}(x)$, represents the code length for any given event x under this scheme. Creating an optimal coding scheme is not necessarily straightforward, and we refer the reader to MacKay's book [208] for details, but we assume that we have one in what follows. Similarly $\mathbf{H}(X \mid Y)$ captures the average code length to encode x given that y occurs, in an optimal encoding scheme for the measurements X given

Y, while $\mathbf{h}(x \mid y)$ represents the code length for any given events x. Then, $\mathbf{I}(X : Y)$ is the average difference in code length between coding the value x in isolation or coding the value x given y, while $\mathbf{i}(x : y)$ represents this difference in such code lengths for any specific events x and y under these schemes. In this way we see that $\mathbf{i}(x : y)$ may be either positive *or negative* for a specific pair x, y. Finally, $\mathbf{I}(X : Y \mid Z)$ is the average difference in code length between coding the value x given z or coding the value x given both y and z. while $\mathbf{i}(x : y \mid z)$ represents this difference in such code lengths for any specific events x, y and z under the optimal schemes.

3.3 Mutual Information and Phase Transitions

So far, we have seen that mutual information has two uses:

1. It tells us if two time sequences share something, a more powerful non-linear measure than correlation. In the case of Gaussian statistics it reduces to correlation.
2. It is the measure of how much information can be transmitted down a noisy channel, the application for which Shannon [304] invented it (Sect. 3.2.7).

But it has at least one other use, the importance of which has soared with the global financial crisis, climate change and numerous other catastrophic or sudden change phenomena. There is a huge body of theory associated with these transitions, and there are numerous books, such as that by Ricard Solé [308]. There is also catastrophe theory from the 1980s, developed by René Thom, but in some ways before its time. Without today's number-crunching power, and particularly computer graphics, it languished in the domain of interesting but abstruse theories. Catastrophe theory in Thom's original description was deterministic.

The other mainstream idea is that of the *phase transition.* The two are not mutually exclusive, and some catastrophes may be loosely described as phase transitions [307], but work on genuine stochastic catastrophe theory is ongoing.

Phase transitions have an *order.* The phase transitions we know about already are the changes in the states of matter – solid to liquid to gas, ice to water to steam. These are *first order,* because at the transition point, 0°C for ice melting or 100°C for water boiling at sea level, there is a discontinuity in some key system variable as a function of a *control parameter,* temperature in the case of ice melting. In the case of the states of water, it is the total energy of the system. With temperature remaining constant, energy is taken in, for the latent heat of ice to water, until all the ice has melted. So going from just below freezing to just above, the energy jumps.

Just to make things confusing, this system variable is often called the *order parameter,* a different use of the word to the order of the transition. A phase transition exhibits a change from order to disorder, or vice versa.

It does not matter too much what this system variable is, and the excitement of these ideas lies in the very wide range of systems, physics, chemistry, biology,

ecology, economics, just to begin with, to which we can apply them. But for this diverse list, it is often *second-order* transitions which are usually of interest.

In a second-order phase transition (PT2), the system variable is continuous, but its first derivative is not. In the physics world, one example of a PT2 which has received a lot of attention is the ferromagnetic–paramagnetic transition in some materials. At a definite temperature, the *Curie point*, the material ceases to be a magnet (i.e. ferromagnetic) at all higher temperatures. A property, the *magnetic susceptibility*, drops to zero at the Curie temperature and stays zero from then on. So, there is a kink at this temperature, which implies the discontinuity in the first derivative.

It turns out that PT2s have several common properties across many systems, be they financial stock markets or frog populations:

- *Increased variance* near the transition.
- *Critical slowing down,* wherein the system takes longer and longer to respond to a small perturbation.
- *Flickering,* where the system briefly flips over to the state on the other side of the phase transition, and then flips back again [297].
- **The MI peaks** at a second-order phase transition across many systems, such as the Ising model [218, 176, 360] and the Vicsek flocking model [353].

It is this last property that is important:

Key Idea 10: *Mutual information peaks at a second-order phase transition, across very many systems.*

Chap. 5 describes the phase transitions in various canonical systems and how the MI peaks accordingly: random Boolean networks and the Ising model. In the latter case, transfer entropy has turned out to be an effective predictor of the phase transition going from the disordered to ordered side (see Sect. 5.2 for details).

3.4 Numerical Challenges

Although Shannon's definition of mutual information dates back to the middle of the twentieth century, it was only at the end that it became feasible to use empirical data. Before then one had to rely on Gaussian or other analytical approximations. There were two reasons for this: the first is that calculating entropy is computationally demanding; the second is that there were many numerical issues, the resolution of which only started to appear in the last decade or so. Simple use of formulae such as Eqn. 3.2, with the probabilities estimated directly, suffers from several problems. In this section we outline some of them and their solutions.

Numerical estimation of entropy measures is a very complex topic, worthy of a book in its own right. So, this section can at best get across the essential ideas, and refer the reader to the more detailed reviews and original papers.

To begin with we look at the primary concept of entropy estimation itself and introduce the ideas of bias and variance. Unfortunately, there is no best estimator for entropy which works across all distributions, thus, with any new problem, some testing of different methods will be necessary. The case of small data sets presents additional challenges, and these have been important in the neuroscience world. We highlight some examples in Sect. 3.4.1.2. Estimation of errors and significance is also non-trivial. Some brief comments appear in Sect. 3.4.1.3.

Now, one might think that, having got good estimators for entropy, mutual information, transfer entropy and their conditional extensions would all follow in a straightforward fashion. Although the theoretical relationships are exact, say, expressing transfer entropy in terms of mutual information, they are not necessarily a good way to approach the estimation from a numerical point of view. Thus, we look at mutual information, and by implication, transfer entropy, as a separate topic in Sect. 3.4.2.

Key Idea 11: *Naively calculating information from frequency estimates is just that, naive!*

3.4.1 Calculating Entropy

Despite its long history, the estimation of entropy, mutual information and transfer entropy is still an ongoing problem, with new analyses and algorithms still appearing. This section attempts to give an overview of some of the issues, but without too much attempt at mathematical rigour in the interests of space.

Fortunately there are good open-source software packages around, which allow the practical researcher interested in applications to steer a reasonably safe course. Bias is intrinsic to entropy estimation, thus, with a little extra bias in the book, we might suggest Barnett's Multivariate Granger Causality Toolbox for Matlab [26], Lizier's Java Information Dynamics Toolkit (JIDT) [183], written in Java (but callable from Matlab, GNU Octave, Python and R) and available from Github. Further information is provided in Sect. 4.3.3, and some examples using these toolkits are included later in this book (e.g. in Chap. 5 and Chap. 7).

3.4.1.1 Plug-in (Max-Likelihood Estimator)

Eqn. 3.2 provides an obvious method for estimating entropic quantities for discrete data by simply estimating the underlying probabilities p from frequencies of occurrence, $\hat{p}_j = \frac{n_j}{N}$ for n_j events in bin j from N samples in total. This simple estimator, referred to as the *plug-in* or *maximum-likelihood estimator* (MLE), is not free of bias

and has a finite variance (defined below). In the discussion which follows, $\mathbf{H}_{MLE}(\hat{p})$ refers to this estimator, where $\hat{p}$ is the plug-in estimate of the true underlying PDF p:

$$\mathbf{H}_{MLE}(\hat{p}) = -\sum_{j=1}^{M} \hat{p}_j \log_2 \hat{p}_j. \tag{3.43}$$

The search goes on for better quality estimators. Paninski [254] provides a thorough review and statistical analysis, and we mention briefly some new developments in Sect. 3.4.1.6. The following results are drawn from Paninski [254], and we refer the reader to his paper for the fine mathematical details, which we do not have the space to include.

The challenges are somewhat different for continuous distributions, and we look at an alternative estimator in Sect. 3.4.1.5, which forms the basis for some of the mutual information estimators in Sect. 3.4.2. A lot of the innovation in entropy estimators comes from neuroscience, from Bialek, de Ruyter and others [318, 255, 240, 239].

First we consider *bias,* i.e. a systematic over-or-under estimation of a quantity. In fact, the entropy is always *underestimated* on average:

$$E\{\mathbf{H}_{MLE}(\hat{p})\} \leq \mathbf{H}(p), \tag{3.44}$$

with the difference between these quantities being the bias B of $\mathbf{H}_{MLE}(\hat{p})$. A bias correction, due to Miller and Madow [223], Eqn. 3.45, has been known for some time [254]:

$$\mathbf{H}_{MM} = \mathbf{H}_{MLE} + \frac{M-1}{2N}, \tag{3.45}$$

where M is the number of bins or symbols and N the number of data points or samples used to construct the PDF.

We can also consider the *variance* across our estimates. Paninski [254] also quotes results for bounds on the variance of $\mathbf{H}_{MLE}$

$$\mathbf{v}(\mathbf{H}_{MLE}) \leq \left(\frac{(\log N)^2}{N} \right) = \mathbf{v}_{max}. \tag{3.46}$$

The errors in the estimation of $\mathbf{H}$ can also be expressed in terms of $\mathbf{v}_{max}$, Eqn. 3.47, as the probability, P, that the error exceeds a threshold, ε, decaying exponentially with ε. Thus, the variance falls almost inversely with N, but the error falls very rapidly with the threshold and N.

$$P(|\mathbf{H}_{MLE} - E\{\mathbf{H}_{MLE}\}|) > \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2\mathbf{v}_{max}}}. \tag{3.47}$$

In general, as the bias B for an estimator improves, the variance gets worse and vice versa, expressed by the ratio, R, Eqn. 3.48, when N is much larger than M

$$R = \frac{\mathbf{v}}{B^2} \approx \frac{N(\log M)^2}{M^2}. \tag{3.48}$$

Since R depends linearly on N but inversely on M^2, the bias will dominate when M is large, except for huge N [254]. When $N \gg M$, this ratio is greater than 1 and the bias corrections are not significant with respect to the variance in the estimates. Note that the Tukey recommendation for the number of bins [233], as $M \approx N^{0.5}$, gives a value of $(\log M)^2$ for R, which is greater than 1. When $R < 1$, the bias corrections become significant.

Key Idea 12: *There is a trade-off between bias and variance in the calculation of entropy.*

3.4.1.2 Estimation for Small Data Sets

Unfortunately, quite a number of situations in which we would like to calculate the entropy are plagued by small data sets. Bonachela et al. [44] point out that there is no perfect entropy estimator: different estimators will perform differently on different data sets. In fact, a bit like the uncertainty principle in quantum mechanics, it is generally impossible to minimise both the variance and the bias.

They introduce the idea of a *balanced* estimator, between bias and variance. Eqn. 3.49 gives the simplest form they derive, but prior knowledge of some characteristics of the probability distribution can improve it.

$$\mathbf{H}^{bal}(X) = \frac{1}{N+2}\sum_{i=1}^{M}\left[(n_i+1)\sum_{j=n_i+2}^{N+2}\frac{1}{j}\right], \tag{3.49}$$

where n_i is the number of occurrences of x in bin i, of which there are M possibilities, and N is the sample size. We can write this using the digamma function, ψ (eqn Eqn. 3.59) as (in *nats*)

$$\mathbf{H}^{bal}(X) = \frac{1}{N+2}\sum_{i=1}^{M}(n_i+1)(\psi(N+3)-\psi(n_i+2)). \tag{3.50}$$

3.4.1.3 Error Estimation

The general variance and bias analysis in Sect. 3.4.1.1 can be made data specific using results obtained by Roulston [289], for the variance in the observed MLE entropy, $\mathbf{H}_{MLE}$, using the above nomenclature:

$$\mathbf{v}(\mathbf{H}_{MLE}(X)) = \frac{1}{N^2}\sum_{i=1}^{M}(\log_2 \hat{p}_i + \mathbf{H}_{MLE}(X))^2\,\mathbf{v}[n_i] \tag{3.51}$$

with

$$\mathbf{v}[n_i] \approx N\hat{p}_i(1-\hat{p}_i) + O(\varepsilon_i), \tag{3.52}$$

where the ε_i are the random errors in the exact probabilities p_i compared with the observed values, $\hat{p}_i = \frac{n_i}{N}$ (as defined earlier):

$$\varepsilon_i = \frac{\hat{p}_i - p_i}{p_i}. \tag{3.53}$$

For the mutual information $\mathbf{I}_{MLE}$ as computed following the MLE technique, the equation is more complicated, reflecting the contribution of several entropy terms

$$\mathbf{v}(\mathbf{I}_{MLE}) = \frac{1}{N^2}\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y}\left(\log_2\left(\sum_{k=1}^{M_X}\hat{p}_{jk}\right) + \log_2\left(\sum_{k=1}^{M_Y}\hat{p}_{ki}\right) + \log_2\hat{p}_{ij} + \mathbf{I}_{MLE}\right)^2 \mathbf{v}[n_{ij}] \tag{3.54}$$

with $\mathbf{v}[n_{ij}]$ defined analogously to Eqn. 3.52:

$$\mathbf{v}[n_{ij}] = N\hat{p}_{ij}(1-\hat{p}_{ij}) + O(\varepsilon_{ij}). \tag{3.55}$$

A practical test for checking the robustness of a mutual information estimator is to randomly permute one of the variables (see further discussion on assessing statistical significance for such measurements, in particular for transfer entropy, in Sect. 4.5.1). The marginal entropies will not change, but the link between the variables is destroyed, so the MI estimation will be distributed as though those variables had no relationship.[8]

3.4.1.4 Kernel Density Estimation

The difficulties of selecting partitions for a direct estimate of the probabilities and subsequently entropies may be circumvented by a statistical technique known as kernel estimation (or kernel *density* estimation for differential entropies; see below). The idea is simple: instead of the bins having fixed rigid boundaries, with a point in one and only one bin, we reconsider or adapt the bins each time we consider the PDF for a given sample. The bins are described by some *kernel function* Θ, which may fall off more gradually at the bin boundaries. Using histograms (as per the previous section), the probability estimate for a value is constructed by locating the bin in which it would fall; the probability is then proportional to the number of samples in the bin, and all other samples are ignored. In contrast, for a kernel density estimate we now include more, possibly all, samples, but the kernel function weights the values according to the distance from the test point, and the weight usually falls monotonically with distance.

Formally (e.g. following [298, 183]), the relevant probability distribution function (e.g. $\hat{p}(x_n)$ for sample n of X) is estimated with a *kernel function* Θ, which

[8] In practical cases, the distribution of estimated MI values here will be above zero – despite the presence of any underlying relationship between the permuted variables – as discussed in Sect. 4.5.1.

measures "similarity" between pairs of samples x_n and $x_{n'}$ using a resolution or *kernel width* r:[9]

$$\hat{p}_r(x_n) = \frac{1}{N}\sum_{n'=1}^{N} \Theta\left(\frac{x_n - x_{n'}}{r}\right). \tag{3.56}$$

A simple choice here is the step kernel $\Theta(|u| \geq 1) = 0$, $\Theta(|u| < 1) = 1$, giving a *box-kernel estimator*. This results in $\hat{p}_r(x_n)$ being the proportion of the N values which fall within r of x_n. This can be thought of as adapting a bin to be centred on our sample x_n with width r. Another common choice is a Gaussian kernel function, which results in a smooth fall-off of the bin boundaries away from the sample x_n; here the kernel radius r would control the fall-off rate (see [231]). Clearly, Eqn. 3.56 can be generalised to multivariates, e.g. by multiplying kernel functions in each dimension.

These plug-in estimates for the PDFs are then used directly in evaluating the Shannon information content (Eqn. 3.1) for each sample $n \in [1,N]$ and averaging these over all samples (Eqn. 3.2) to obtain the entropy, i.e.

$$\mathbf{H}(X) = -\frac{1}{N}\sum_{n=1}^{N} \log_2 \hat{p}_r(x_n). \tag{3.57}$$

Let us be clear that the definition Eqn. 3.56 provides an adaptive *discrete* (or bin/histogram-based) probability distribution function estimate $\hat{p}_r(x_n)$, and so will result in producing a kernel estimator for Shannon entropy.

In contrast, if we wish to compute a differential entropy (see Sect. 3.2.5), then one should correct Eqn. 3.56 to form a probability *density* function (e.g. see Sect. 2.5.3 and [231, 149]). Generally this means dividing Eqn. 3.56 by a factor of r, or r^d for multivariate kernel functions, where d is the number of dimensions of the multivariate space. Furthermore, it is crucial then that the kernel function be a valid probability density function itself, i.e. integrate to 1 [231, 149]. This produces a kernel *density* estimator for differential entropy. That is, one would use $\hat{p}_r(x_n)/r$ to correct the density for the spatial scale in our above example, but also alter $\Theta(|u| < 1) = \frac{1}{2}$ for proper normalisation.[10] In any case, for mutual information calculations (and conditional MIs, e.g. transfer entropy), these correction factors on each PDF cancel (assuming the same kernel width is used for each variable), meaning that either approach may be used.

Kernel estimation can measure non-linear relationships and is model-free, though it is sensitive to the parameter choice for resolution r [298, 149]. Selecting a value for r can be difficult, with too small a value yielding under-sampling effects while too large values ignore subtleties in the data. One can heuristically determine a lower

[9] Though r is more properly a kernel radius rather than a width.

[10] Intuitively, we can simply think of this as dividing by $2r$ to correct $\hat{p}_r(x_n)$ for the $2r$ space the neighbours were counted over.

bound for r to avoid under-sampling (see [183]). Finally, we note that the estimates provided by kernel estimation contain a bias; methods for bias correction here are available for individual entropy estimates (e.g. see [118] for the box kernel).

3.4.1.5 Continuous Distributions: Digamma and Kozachenko–Leonenko Estimators

An alternate way of estimating the differential entropy of a continuous variable from a finite number of samples looks nothing like the canonical formulae. Suppose we have N (univariate) samples, x_i, and we order them in increasing size, as $x_1 < x_2 \ldots < x_N$. Then, the digamma entropy estimator is given by (see e.g. [168])

$$\mathbf{H}(X) \approx \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i) - \psi(1) + \psi(N) \tag{3.58}$$

in *nats*, where ψ is the digamma function, which is effectively the derivative of the log of the gamma function $\Gamma(x)$, i.e.

$$\psi(K) = \frac{1}{\Gamma(K)} \frac{d\Gamma(K)}{dK}. \tag{3.59}$$

The digamma function can be defined recursively as

$$\psi(K+1) = \psi(K) + \frac{1}{K} \tag{3.60}$$

with $\psi(1) = -C$ given in terms of the Euler constant, C, also called the Euler–Mascheroni constant:

$$C = 0.5772156\ldots \tag{3.61}$$

Kozachenko and Leonenko [166] then extend the estimator in Eqn. 3.58 in that the distances between sorted neighbouring points $x_{i+1} - x_i$ are replaced by Kth nearest-neighbour distances in d-dimensional space, $\frac{\varepsilon_i}{2}$, using their notation (in Eqn. 3.62). The division by two appears because we want to look at bands in the marginal planes of width twice the nearest-neighbour distance. The Kozachenko–Leonenko estimator is given by

$$\mathbf{H}_{nn}(X) = \psi(N) - \psi(K) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \varepsilon_i \tag{3.62}$$

in *nats*, where d is the dimension of x and c_d is the volume of the d-dimensional unit ball. Thus, for the maximum norm, $\log c_d$ vanishes. Comparing Eqn. 3.62 for $K = 1$ with Eqn. 3.58 for large N, they are roughly the same, since $x_{i+1} - x_i$ or $x_i - x_{i-1}$ is

the nearest-neighbour distance for x_i. Understandably, we refer to this estimator as a member of the class of "nearest-neighbour" estimators.

3.4.1.6 The State of the Art

The unfortunate fact that there is no perfect estimator for entropy fosters ongoing research into ever more finely tuned methods. Since the estimators depend on the data statistics, Bayesian methods with some prior assumptions about the statistics can improve on the estimators we have discussed. An important innovation from Nemenman et al. [240] is to change the way one models the initial assumptions about what the probability distribution might be. It turns out that assuming a prior distribution of the *entropies* as opposed to the probability distribution can give significantly better results. One such recent analysis by Vinck et al. [339] develops such estimators further, with some improvements at the expense of considerable computational complexity.

3.4.2 Calculating Mutual Information

When we come to estimate mutual information, we find a range of different contexts to consider. To begin with we have systems where we know the dynamical behaviour and have an explicit, continuous, model. Common examples are chaotic systems, such as the Rössler, used frequently as a canonical test for MI algorithms. In these systems we can choose when and where to take the sample points we need. The other systems are typical of experiment or observation, where we have a set of data points, we do not have much control over when they are taken (such as daily stock market data (Sect. 6.2.3)), and do not always have a theoretical model.

We can then subdivide the algorithms into two categories:

1. Methods dealing with already discrete or otherwise binned (discretised) data, where the calculation is based on the probability functions above with variations on the way the sample points are grouped together into bins.
2. Methods operating on continuous data which use some indirect approach, such as the nearest-neighbour method we have already seen for entropy (Sect. 3.4.1.5).

The simplest estimators for mutual information are simply extensions of the algorithms of the above two types for entropy, combining estimations of the joint and marginal entropies as per Eqn. 3.10. For example, Moon et al. [231] introduced kernel density estimators for the calculation of mutual information, with a view to getting better estimates from small data sets. Yet such extensions may not work as well as one hopes; for example bias correction for individual kernel estimates of entropy [118] is not directly generalisable to sums of entropies because "the finite sample fluctuations... are not independent and we cannot correct their bias separately" [149].

Improvements can be made however by tailoring algorithms specifically to the mutual information rather than the underlying entropies alone. We consider such algorithms in this section. A good starting point for the reader wanting detailed knowledge of the numerical procedures is the article by Cellucci et al. [55], which covers the issues of partitioning deeply and introduces the simple but effective algorithm we discuss in Sect. 3.4.2.1.

Key Idea 13: *Calculating mutual information is tricky and needs to be validated case by case.*

3.4.2.1 Bins, Fixed and Adaptive

In many practical cases, we just have a set of samples or points, collected somehow. Where these are discrete-valued samples, we can proceed directly to plug-estimates on these "bins" and improvements on these techniques in Sect. 3.4.1.1 and Sect. 3.4.1.2. Otherwise, where we have continuous-valued samples, then to use a binning-based approach, the first step is to put these samples into bins to determine the probability distributions. If we make the bins too small, then the values will be very noisy, and in the limit of one point per bin, the distribution will be meaningless. If the bins are too big, then the precision in the PDF will be very low. Statisticians have already sorted this out for us [55]: an early, rough estimate is $N^{\frac{1}{2}}$ from Tukey [233], but there are better, albeit more difficult to calculate, measures now available. Slonim et al. [306] offer the alternative of dynamically testing multiple subdivisions and provides criteria for selecting the best one.

But we are still not out of the woods. A uniform partitioning in X and Y, is not optimal. The procedure introduced by Cellucci et al. [55] is to use *partitions of varying size,* such that the number of samples in each marginal bin along the X and Y axes is equal (with the same number of bins $M_E = M_X = M_Y$ along each axis).[11] If there are N samples then the number of bins M_E for each marginal is chosen to be the largest integer such that $\frac{N}{M_E^2} \geq 5$; in other words, such that there are at least five samples on average in each of the M_E^2 bins in the joint X–Y space. The marginal entropies are now simply given by $\log_2 M_E$ (for N a multiple of M_E), and using this partitioning it is now possible to calculate $p(X,Y)$, hence the joint entropy and the mutual information via Eqn. 3.10. The search for better mutual information estimators for discrete variables continues. The surge in big data has driven the need for better estimators of shared information and causality, sometimes now with large numbers of data categories. Seok and Kang [300] introduced a new partitioning

[11] Since an equal number of samples in each bin will give the maximum possible entropy for the marginal distributions over X and Y, this is sometimes referred to as a *maximum entropy binning*. Obviously having precisely the same number of samples is only possible where N is an integer multiple of M_E.

algorithm for mutual information calculation, which shows big improvements on simulated data. Similar to the ideas presented in Cellucci et al. [55], the goal is to find partitions in the marginal densities, where the probabilities are uniform. But now sub-categorisation of each variable is sought, in which each sub-category is uniform. The joint distribution is then recursively made uniform. Early results for health care data of several thousand records look promising.

3.4.2.2 The KSG (Kraskov, Stögbauer and Grassberger) Algorithm

Kraskov et al. [168] created an effective estimator for MI for continuous distributions by extending entropy estimators in Sect. 3.4.1.5. Rather than attempt to estimate the probability distributions directly (as the kernel estimator in Sect. 3.4.1.4 does), it makes use of nearest-neighbours. Getting an intuition for this estimator is even more difficult than the digamma estimator for the entropy!

There are two slightly different algorithms to estimate $\mathbf{I}(X:Y)$. Both start with a variation on the Kozachenko–Leonenko estimator $\mathbf{H}_{nn}(X)$ [166] (see Eqn. 3.62).

Now we sit on one point in the joint space, $z_i = \{x_i, y_i\}$, and measure the distance, ε_i, to the K^{th} nearest neighbour. The distance metric is the maximum norm:

$$||z - z_i|| = \max(||x - x_i||, ||y - y_i||). \tag{3.63}$$

The x and y norms do *not* have to be maximum norms, and they could be quite different spaces, with quite different norms. So, we might want to find out how much information there is between the amount of sunshine at the beach (hours) and the amount of ice cream sold (litres, tons, maybe, not in England, though).

Now, imagine that x and y have no mutual information, that we have randomly scattered values. Then for any given y, x could be anything. Suppose we consider the mutual information between something unlikely to affect ice cream sales, say, the price of modern art. When we discuss transfer entropy, we will consider the importance of conditioning out other factors (Sect. 4.2.3). It is possible to think up all sorts of strange indirect links, but let us assume here that there is no direct link. Thus, for any given art price, the probability of selling particular amounts of ice cream will be the same. If we take a sample of ice cream volume and art price (maybe measured at random times) and plot them on a two grid, they will be scattered randomly over it.

For the case of sunshine, however, the situation is quite different. High sunshine means a cluster of ice cream sales around high volume.

The first algorithm, given by Eqn. 3.64, then works like this: For each point we find its K^{th} nearest neighbour in the joint X–Y space, a distance $\frac{\varepsilon}{2}$ away (in the notation of the $\mathbf{H}_{nn}(X)$ estimator in Eqn. 3.62). Since this is the maximum norm over those taken in the X and Y spaces, we project this value back onto the X and Y axes and count the number of points, n_x and n_y, *strictly within* a row (X) and column (Y) of width ε.

For multi-dimensional X and Y these rows and columns become hyper-stripes.

$$\mathbf{I}^{(1)} = \psi(K) + \psi(N) - E\left\{\psi(n_x+1) + \psi(n_y+1)\right\}. \tag{3.64}$$

Key Idea 14: *The key innovation of the KSG algorithm is getting the numerical errors to partially cancel in the marginal and joint entropy estimates.*

It does this by finding the nearest-neighbour distance for the joint distribution $\{X,Y\}$. Then it works backwards to find what the number of nearest neighbours would have been for each of the marginal distributions, in order for that to have been the nearest-neighbour distance. So the K value is different for the marginals, and is obtained from the nearest-neighbour counts. This is valid because $\mathbf{H}_{nn}(X)$ in Eqn. 3.62 holds for any value of K and does not have to be fixed when estimating the marginal entropies. So, combining $\mathbf{H}_{nn}(X)$, $\mathbf{H}_{nn}(Y)$ and $\mathbf{H}_{nn}(X,Y)$ for a fixed value of K in the $\{X,Y\}$ space keeps the same scale in all spaces, with bias terms of the same order that oppose each other and (mostly) cancel.

To get a bit more intuition for this estimator, let us go back to ice cream and modern art. Since the distribution is random, let us say across a square of size D, there will be a density $\rho = \frac{N}{D^2}$ of points per unit area. So the column and row will have around $D\varepsilon\rho = \frac{N\varepsilon}{D}$ points. The number of nearest neighbours, K, determines the value of ε as approximately $D\sqrt{\frac{K}{N}}$. Making use of the asymptotic approximation $\psi(x) \approx \log(x)$, then

$$E\left\{\psi(n_x+1) + \psi(n_y+1)\right\} \approx \log(KN) \approx \psi(K) + \psi(N). \tag{3.65}$$

with a generous neglect of smallish terms. Thus, $I \approx 0$ as expected.

The maximum mutual information occurs when one variable completely predicts the other. So, imagine, then, that our sample points form a tight cluster around a diagonal line. Then row and column bands miss most of the points. Thus, n_x and n_y are small ($\rightarrow K$) and I is maximal.

As this line swings round to vertical, the y value strongly predicts the x value because it is always the same. But x has no predictive value ($n_x \rightarrow N$), and mutual information is symmetric. Its value in this case is close to zero, because the entropy of x is very low. Similarly $\mathbf{H}(X|Y)$ is very low, because there is simply no entropy in X regardless of Y. Thus, we can see from Eqn. 3.9 that the mutual information will be very small.

The extension to the second algorithm involves reinterpretation of n_x and n_y. Instead of using the same ε for both x and y, we use the respective x,y norms $\varepsilon_x, \varepsilon_y$ from the K^{th} nearest neighbour for the width of the column and row about each point. We then count points within *or on* these boundaries, and estimate the MI as

$$\mathbf{I}^{(2)} = \psi(K) + \psi(N) - \frac{1}{K} - E\left\{\psi(n_x) + \psi(n_y)\right\}. \tag{3.66}$$

Keeping in mind that entropy estimators are PDF dependent, estimator 1 is likely to have a lower statistical but higher systematic error than 2. The second estimator is likely to prove better when the dimensionality is high. Similarly, the best values of K need to be determined empirically, though generally the estimator is robust to the choice of K from $k = 4$ upwards, as variance in the estimate decreases with K.

3.4.2.3 KSG Estimator for Conditional Mutual Information

As we shall see in the next chapter (Sect. 4.3.1), to calculate the transfer entropy we effectively need to estimate the conditional mutual information. To achieve the same cancellation of errors, we need to go back to the original entropies and apply the same argument for back-estimating the number of nearest neighbours.

Starting with Eqn. 3.16, we can rewrite the conditional entropies in terms of joint entropies as follows:

$$\mathbf{I}(X : Y \mid Z) = \mathbf{H}(X,Z) + \mathbf{H}(Y,Z) - \mathbf{H}(X,Y,Z) - \mathbf{H}(Z)\,. \tag{3.67}$$

Now, we calculate the nearest-neighbour distance for the three-variable, joint distribution, and calculate the effective number of neighbours for each marginal. For the two-variable joint terms, such as $\mathbf{H}(X,Z)$, we have to count the number of neighbours in a box surrounding each point, as opposed to a stripe. This is explained in detail in Sect. 4.3.1.

3.4.3 The Non-stationary Case

The assumption underlying all these measures based on time series is that, over the window for which the probabilities are calculated, the series is stationary. Unfortunately, this is not always the case, and misleading results may occur. The estimate of mutual information is not meaningful in this case [340]. Nason [236] presents new, powerful methods, based on wavelets, for determining stationarity, to which we refer the interested reader.

Open Research Question 2: *Can wavelet methods be used to get better mutual information for non-stationary systems?*

Furthermore, approaches utilising an ensemble of repeated trials in neuroscientific experiments are discussed in Sect. 4.3.1.1.

Chapter 4
Transfer Entropy

In this chapter we get to the essential mathematics of the book—a detailed discussion of transfer entropy. To begin with we look at the basic formalism (Sect. 4.2) and some variants thereof, which appear in later chapters (Sect. 4.2.5). We then go on to compare it with the earlier, closely related concept of *Granger causality* (Sect. 4.4). The relevance to phase transitions is taken up in Sect. 4.6, and the chapter concludes with extension of the discrete-time case to continuous-time processes (Sect. 4.7).

4.1 Introduction

Given jointly distributed random variables X, Y—discrete or continuous, and possibly multivariate—we have seen in Chap. 3 that the mutual information $\mathbf{I}(X : Y)$ furnishes a principled and intuitive answer to the questions:

- How much uncertainty about the state of Y is resolved by knowing the state of X (and vice versa)?
- How much information is shared between X and Y?
- How may we quantify the degree of statistical dependence between X and Y?

Suppose now that, rather than *static* variables, we have jointly distributed *sequences* of random variables X_t, Y_t labelled by a sequentially enumerable index $t = \ldots, 1, 2, 3, \ldots$. Intuitively the processes X_t, Y_t may be thought of as an evolution in *time* (t) of some unpredictable variables X, Y, that is, *random time-series processes* (Sect. 2.3.5). Such joint or multivariate *stochastic processes* are natural models for a huge variety of real-world phenomena, from stock market prices to schooling fish to neural signals, which may be viewed (generally through lack of detailed knowledge) as non-deterministic *dynamic* processes.

How, then, might we want to frame, interpret and answer comparable questions to the above for dynamic stochastic processes rather than static variables? We may, of course, consider the mutual information $\mathbf{I}(X_t : Y_t)$ between variables at a given fixed time t. But note that, by *jointly distributed* for stochastic processes, we mean that

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_4

there may be dependencies within any subset $\{X_t, Y_s : t \in T, s \in S\}$ of the individual variables. Thus, for instance, X_t, the variable X as observed at time t, may have a statistical dependency on its value X_{t-s} at the earlier time $t-s$, or indeed on its entire *history* $X_{t-1}, X_{t-2}, \ldots$, or the history $Y_{t-1}, Y_{t-2}, \ldots$ of the variable Y. A particularly attractive notion is that of quantifying a time-directed *transfer* or *flow* of information between variables. Thus we might seek to answer the question:

- How much information is transferred (at time step t) from the past of Y to the current state of X (and vice versa)?

This information transfer, which we would expect—unlike the contemporaneous mutual information $\mathbf{I}(X_t : Y_t)$—to be asymmetric in X and Y, is precisely the notion that transfer entropy aspires to quantify.

4.2 Definition of Transfer Entropy

The notion of *transfer entropy* (TE) was formalised by Thomas Schreiber [298] and independently by Milan Paluš [253],[1] although it may be argued that, historically, similar concepts have surfaced periodically in various guises since as early as the 1950s [354], partly via a somewhat tangled shared provenance with the closely related concept of *Wiener–Granger causality* [354, 112, 114, 105, 285] (See Sect. 4.2.4.) Amblard [2] provides a useful historical review of this area.

Schreiber and Paluš realised that an obvious candidate for a time-asymmetric measure of information transfer from Y to X, namely the *lagged* mutual information $\mathbf{I}(X_t : Y_{t-s})$ [298, 149], is unsatisfactory for the reason that it fails to take into account *shared history* (as well as common external driving influences) between the processes X and Y, and that this is likely to lead to spurious inferences of directed information transfer. This is neatly illustrated by a minimal example (4.1), which we adapt from [149].

Example 4.1. In this example X_t, Y_t is a two-variable, first-order, stationary Markov chain (Sect. 2.3.6) with X and Y binary variables taking values ± 1. The time index runs from $t = -\infty$ to $t = +\infty$. The process Y is autonomous, in the sense that its current state depends only on its own past and has no dependency at all on X. It transitions deterministically from state y to state $-y$ (i.e. it flips) at each successive time step. The current state of X, on the other hand, has no direct dependence on its own history, but depends probabilistically on the state of Y at the previous time step. Specifically, at each time t, $X_t = Y_{t-1}$ with probability $\frac{1+c}{2}$ and $X_t = -Y_{t-1}$ with probability $\frac{1-c}{2}$ for some constant $-1 \leq c \leq 1$. The joint transition probabilities are thus

[1] On an historical note, the term "transfer entropy" was coined by Schreiber in [298], which preceded Paluš' publication [253] by a few months. As is oft the way in science, it is thus Schreiber's formalism that is the more influential and most often cited, although Paluš' exposition is somewhat more general and purely information theoretic in spirit. Paluš also makes explicit the link with Granger causality.

$$\mathbf{P}\big(X_t = x', Y_t = y' \,\big|\, X_{t-1} = x, Y_{t-1} = y\big) = \\ \delta(y', -y)\left[\delta(x', y)\frac{1+c}{2} + \delta(x', -y)\frac{1-c}{2}\right]. \quad (4.1)$$

Stationarity implies that the distribution of the joint process at any time t is given by

$$\mathbf{P}(X_t = x, Y_t = y) = \frac{1}{2}\left[\delta(x, -y)\frac{1+c}{2} + \delta(x, y)\frac{1-c}{2}\right]. \quad (4.2)$$

The marginal probabilities are $\mathbf{P}(X_t = x) = \mathbf{P}(Y_t = y) = \frac{1}{2}$.

Intuitively, a useful measure of information transfer should yield zero in the $X \to Y$ direction (since Y is autonomous), while we might expect to see a non-zero transfer of information in the $Y \to X$ direction (since X depends on the past state of Y). But from (4.1) and (4.2) we have $\mathbf{P}(X_t = x, Y_{t-1} = y) = \mathbf{P}(Y_t = y, X_{t-1} = x) = \frac{1}{2}[\delta(x, y)\frac{1+c}{2} + \delta(x, -y)\frac{1-c}{2}]$, and we may calculate, working to a single *lag*—that is, a single step back in time (*cf.* [149]):

$$\mathbf{I}(X_t : Y_{t-1}) = \mathbf{I}(Y_t : X_{t-1}) = \tfrac{1}{2}\left[(1+c)\log(1+c) + (1-c)\log(1-c)\right]. \quad (4.3)$$

Since (at least if $c \neq 0$) $\mathbf{I}(Y_t : X_{t-1}) > 0$, lagged mutual information, as a notional measure of information transfer, suggests a spurious transfer of information in the $X \to Y$ direction. The explanation for this failure is that there is indeed shared information between the previous state of X and the current state of Y: in this case (if $c \neq 0$), knowing X_{t-1} tells us something about Y_{t-2} which, in turn, tells us something (in fact everything!) about Y_t; in other words, X and Y *share a common history*.

The problem in the above example is essentially that, even without explicit knowledge of the past of Y, the past of X already yields information about its own current state.

Key Idea 15: *Schreiber and Paluš' insight was that, to assess the influence of the past of Y on current X, the shared information between X and its own past must be accounted for.*

Information theory supplies just the tool to effect this accounting: we must *condition on* the past of X as a conditional mutual information (Sect. 3.2.3, Eqn. 3.16 and Eqn. 4.4). Such conditioning removes any redundant or shared information between current X and its own past, but also includes any synergistic information about current X in the source Y that can only be revealed in the context of the past of X.[2]

This motivates the definition of transfer entropy for the special case of history length (lag) 1:

[2] Williams and Beer [359] continue this partial information decomposition of the TE to label the synergistic component as *state-dependent* transfer entropy and the unique component from the source as *state-independent* transfer entropy. See also [195, 28].

Definition 4.1.

$$\begin{aligned}\mathbf{T}_{Y\to X}(t) &\equiv \mathbf{I}(X_t : Y_{t-1} \mid X_{t-1}) \\ &= \mathbf{H}(X_t \mid X_{t-1}) - \mathbf{H}(X_t \mid X_{t-1}, Y_{t-1}).\end{aligned} \tag{4.4}$$

We can then say that:[3]

Key Idea 16: $\mathbf{T}_{Y\to X}(t)$ *with lag 1 may be interpreted intuitively as the degree of uncertainty about current X resolved by past Y and X, over and above the degree of uncertainty about current X already resolved by its* own *past alone.*

Note that $\mathbf{T}_{Y\to X}(t)$, as a conditional mutual information, is always *non-negative* (inclusion of Y_{t-1} in the conditioning variables cannot increase the conditional entropy). Previously (Sect. 3.2.2), we have also seen that mutual information may be interpreted as a measure of *statistical dependence*. Thus we have an intuitive interpretation for vanishing $\mathbf{T}_{Y\to X}(t)$:

$\mathbf{T}_{Y\to X} = 0 \iff$ *X, conditional on its own past, is independent of the past of Y.*

We shall generally refer to X as the *target* and Y as the *source* variable. Note that we retain the argument t in the definition of $\mathbf{T}_{Y\to X}$; although the process in the example above was stationary (and [298] only considered stationary processes), Definition 4.1 makes sense equally for *non*-stationary processes, in which case the PDFs and therefore the transfer entropy will generally depend on the time t. Estimation of transfer entropy from non-stationary empirical time-series data, however, will require some special techniques (Sect. 4.3.1.1); otherwise spurious results may be obtained [340]. For stationary processes we omit the time argument.

Returning to Example 4.1, we may calculate

$$\mathbf{H}(X_t \mid X_{t-1}, Y_{t-1}) = \log 2 - \tfrac{1}{2}\left[(1+c)\log(1+c) + (1-c)\log(1-c)\right], \tag{4.5}$$

$$\mathbf{H}(X_t \mid X_{t-1}) = \log 2 - \tfrac{1}{2}\left[(1+c^2)\log(1+c^2) + (1-c^2)\log(1-c^2)\right], \tag{4.6}$$

$$\mathbf{H}(Y_t \mid X_{t-1}, Y_{t-1}) = \mathbf{H}(Y_t \mid Y_{t-1}) = 0, \tag{4.7}$$

[note that (4.7) holds since Y transitions autonomously and deterministically] so that (*cf.* [149])

$$\begin{aligned}\mathbf{T}_{Y\to X} &= \tfrac{1}{2}\left[(1+c)\log(1+c) + (1-c)\log(1-c)\right] \\ &\quad - \tfrac{1}{2}\left[(1+c^2)\log(1+c^2) + (1-c^2)\log(1-c^2)\right],\end{aligned} \tag{4.8}$$

$$\mathbf{T}_{X\to Y} = 0, \tag{4.9}$$

[3] We remark that this is closer to the approach of Paluš [253]. Schreiber [298] derived his formulation in somewhat different terms—see below for details.

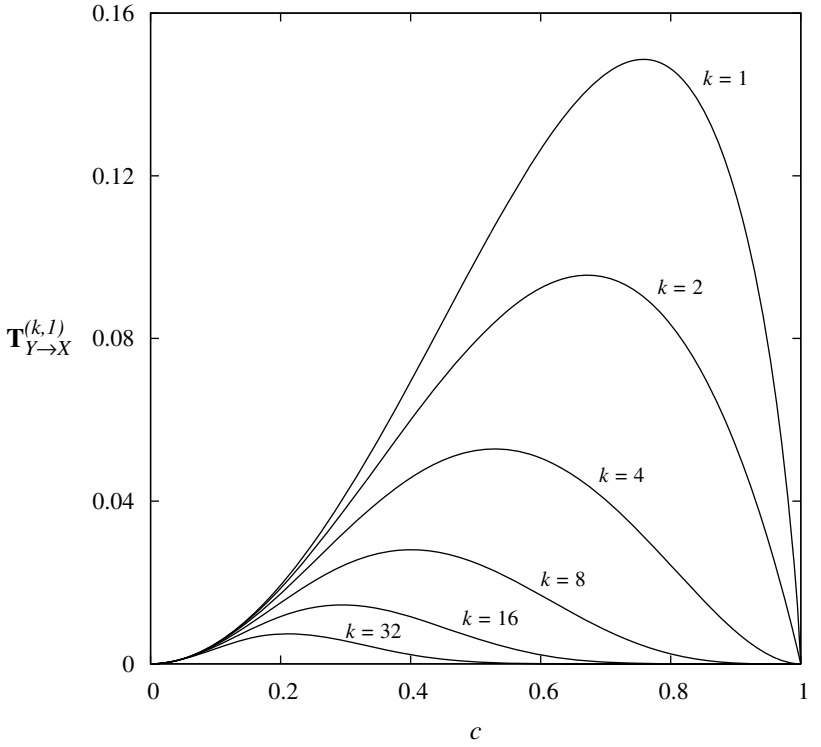


Fig. 4.1 Transfer entropy $\mathbf{T}^{(k,1)}_{Y\to X} = 0$ plotted against coupling parameter c for increasing target history length k for Example 4.1

and therefore the transfer entropy correctly yields non-zero information transfer in the $Y \to X$, but not in the $X \to Y$, direction (see Fig. 4.1, $k = 1$ plot).

4.2.1 Determination of History Lengths

So far we have only considered histories of length 1 for both target and source variables. But what if the shared information between the target and its past extends to a longer history length? What if the earlier values of the source contain additional information about the target? How, then, should we specify history lengths in general? Broadly speaking, too short a history for the target variable risks over-estimating transfer entropy, since we may fail to condition out the full influence of the past of the target on itself. Too long a history for the target variable also risks over-estimating transfer entropy due to under-sampling the multi-dimensional PDFs. Conversely, too short a history for the source variable risks under-estimating

transfer entropy, since we may fail to incorporate the full influence of the past of the source on the target variable. (Note, though, that this simplified argument ignores the possible effects of *synergy* between historical states of target and source variables on the current state of the source variable—see Sect. 3.2.3.1, Sect. 4.2 and [195, 28] regarding synergy in TE.)

4.2.1.1 Target History Length

Schreiber [298] mainly addresses the case where the target variable is a kth-order Markov process (Sect. 2.3.6), and specifies a history length of k for the target variable in the expression for transfer entropy. This ensures that the entire (independent) historical influence of the target variable on its current state is conditioned out.

To frame this mathematically, we introduce the notation

$$\boldsymbol{U}_t^{(k)} \equiv (U_t, U_{t-1}, \ldots, U_{t-k+1}). \tag{4.10}$$

for the length-k history of a variable U, up to and including time t. We note that this is a Takens embedding vector of embedding dimension k and embedding delay $\tau = 1$ (see Sect. 2.3.5). Recall formally that the underlying *state* of a Markov process U is captured by a sufficiently embedded vector $\boldsymbol{U}_t^{(k)}$ (i.e. where the embedding length is greater than the order of the Markov process). This means that, once we move to examine proper embeddings of the time series, the transfer entropy considers *state transitions* of the target $\boldsymbol{X}_t^{(k)} \rightarrow \{X_{t+1}, \boldsymbol{X}_t^{(k)}\}$, and

Key Idea 17: *Transfer entropy measures how much information the source process provides about state transitions in the target.*

Of course, one could also use an embedding delay $\tau > 1$ should this produce more appropriate embedding vectors $\boldsymbol{U}_t^{(k,\tau)}$ (see Sect. 2.3.5). Similarly, it is possible to compute TE with *non-uniform embeddings* of the variables, i.e. selecting *k irregularly spaced* variables U_{t-i} as a representation of the past state of U (see [85]). For simplicity in this book however, we will concentrate on representing TE with standard embedding vectors, with embedding delay $\tau = 1$.

Schreiber also suggests that, if the target variable is *non*-Markov, we should let its history length $k \rightarrow \infty$ (see further discussion in [195]).

Note, however, that even if the *joint* process X_t, Y_t is Markov, the marginal (target and source) variables X_t and Y_t will generally *not* be Markov processes; for example, values in the past of X may include information about X_t that is redundant with that held by Y, even for X_{t-m} for m beyond the joint Markovian order. This is the case, for instance, in Example 4.1, where the joint process is first-order Markov, as is the Y_t process, but (as may easily be verified) X_t is not Markov. We consider this in more

detail in Sect. 4.2.2, recommending that, for proper interpretation as information transfer, the target should be embedded before source embedding is considered.

4.2.1.2 Source History Length

It is less clear how much history of the source variable should be taken into account. If the target variable is kth-order Markov, Schreiber suggests a history length of k or 1 for the source variable, although if the *joint* process is Markov of order ℓ (or, less stringently, if X_t is known to only depend on ℓ lags of Y_t) it would seem to make more sense to take a history length of ℓ for the source variable (taking a longer history will not alter the result in this case); see further discussion in [184]. In one sense, history length for the source variable is an open choice; notwithstanding, we take the view that there is no harm in (theoretically) taking infinite histories for both source and target variables: this ensures that all relevant history is always accounted for.

4.2.1.3 Empirical Determination of History Lengths

Of course, for empirical estimation of transfer entropies from finite time series (Sect. 4.3), practical choices of history lengths will be severely constrained by the amount of data available (the data requirement of transfer entropy scales exponentially with history length), and some scheme for truncating histories will be required. For example, Wibral et al. [350] suggest using the Ragwitz and Kantz criterion [279] of setting the history or embedding lengths (and embedding delays, if these are used also) to provide minimal error in predicting the next value of each series.[4]

4.2.1.4 General (k,ℓ)-History Definition of Transfer Entropy

We can now define the general form of the (k,ℓ)-history transfer entropy:

Definition 4.2.

$$\begin{aligned}\mathbf{T}^{(k,\ell)}_{Y\to X}(t) &\equiv \mathbf{I}\left(X_t : \boldsymbol{Y}^{(\ell)}_{t-1} \,\middle|\, \boldsymbol{X}^{(k)}_{t-1}\right)\\ &= \mathbf{H}\left(X_t \,|\, \boldsymbol{X}^{(k)}_{t-1}\right) - \mathbf{H}\left(X_t \,|\, \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}^{(\ell)}_{t-1}\right).\end{aligned} \tag{4.11}$$

Key Idea 18: $\mathbf{T}^{(k,\ell)}_{Y\to X}(t)$ *may be interpreted intuitively as the degree of uncertainty about current* X *resolved by the past* states Y *and* X*, over and above the*

[4] Note that this criterion does not account for synergies between the source and target's pasts, and so should be extended in the future.

degree of uncertainty about current X already resolved by its own *past* state *alone.*

$\mathbf{T}^{(\infty,\ell)}_{Y\to X}(t)$, $\mathbf{T}^{(\infty,\infty)}_{Y\to X}(t)$ etc. denote the corresponding limits (if they exist). Again $\mathbf{T}^{(k,\ell)}_{Y\to X}(t)$ is non-negative, and for stationary processes we drop the time dependency argument t. If history lengths are clear or irrelevant we also omit the superscripts.

In Schreiber's original formulation [298] he in fact defines $\mathbf{T}^{(k,\ell)}_{Y\to X}(t)$ for a Markov process X_t, equivalently, as the KL divergence (Sect. 3.2.4) between the distributions of X_t conditional on just $\boldsymbol{X}^{(k)}_{t-1}$, and on both $\boldsymbol{X}^{(k)}_{t-1}$ and $\boldsymbol{Y}^{(\ell)}_{t-1}$, yielding the alternative formula

$$\mathbf{T}^{(k,\ell)}_{Y\to X}(t) = \sum_{x_t,\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}} p\left(x_t,\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}\right)\log_2\frac{p\left(x_t\,\middle|\,\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}\right)}{p\left(x_t\,\middle|\,\boldsymbol{x}^{(k)}_{t-1}\right)} \tag{4.12}$$

$$= \sum_{\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}} p\left(\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}\right)\sum_{x_t} p\left(x_t\,\middle|\,\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}\right)\log_2\frac{p\left(x_t\,\middle|\,\boldsymbol{x}^{(k)}_{t-1},\boldsymbol{y}^{(\ell)}_{t-1}\right)}{p\left(x_t\,\middle|\,\boldsymbol{x}^{(k)}_{t-1}\right)}, \tag{4.13}$$

where $p(\cdot), p(\cdot|\cdot)$ denote the (conditional) probabilities of corresponding (histories of) states.

4.2.2 Computational Interpretation as Information Transfer

The conditioning on the past history $\boldsymbol{X}^{(k)}_{t-1}$ of the target X plays an important role in giving TE an interpretation in terms of *distributed information processing* [188, 182].

In the first place, Schreiber's original description of TE [298] can be rephrased as information provided by the source about a *state transition* $\boldsymbol{x}^{(k)}_t \to x_{t+1}$ in the target (or including redundant information $\boldsymbol{x}^{(k)}_t \to \boldsymbol{x}^{(k)}_{t+1}$). This is seen in that the $\boldsymbol{x}^{(k)}_t$ are *embedding vectors* [320] capturing the underlying *state* of the process X for Markov processes (see Sect. 4.2.1.1). We can then consider TE in the wider context of where information is contributed for this state transition or *computation* of the next value X_t. A first step there is to examine how much information is contained in the past state $\boldsymbol{X}^{(k)}_{t-1}$ of X about its next value X_t, the *active information storage* (AIS) [198]:

Definition 4.3.

$$\mathbf{A}_X^{(k)}(t) \equiv \mathbf{I}\left(\boldsymbol{X}_{t-1}^{(k)} : X_t\right) \tag{4.14}$$

$$= \mathbf{H}(X_t) - \mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}\right). \tag{4.15}$$

The *entropy rate* term $\mathbf{H}'_X(t) = \mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}\right)$ includes any information transferred from other variables to X, plus any remaining intrinsic uncertainty. Expanding $\mathbf{H}'_X(t)$ we see that AIS is complementary to the transfer entropy terms, since they are non-overlapping components of the information in X_t [196]:

$$\mathbf{H}(X_t) = \mathbf{I}\left(\boldsymbol{X}_{t-1}^{(k)} : X_t\right) + \mathbf{I}\left(X_t : \boldsymbol{Y}_{t-1}^{(\ell)} \,\middle|\, \boldsymbol{X}_{t-1}^{(k)}\right) + \mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}\right) \tag{4.16}$$

$$= \mathbf{A}_X^{(k)}(t) + \mathbf{T}_{Y\to X}^{(k,\ell)}(t) + \mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}\right). \tag{4.17}$$

The above equations demonstrate how the computation of the next value X_t is composed of stored information and transferred information from $\boldsymbol{Y}_{t-1}^{(\ell)}$,[5] and the consideration of the past history $\boldsymbol{X}_{t-1}^{(k)}$ serves to separate the two. Indeed, it is clear that using too short a history length k will serve to under-estimate the AIS, perhaps over-estimating the TE by confusing some stored information as having been transferred.

Finally, recall from Sect. 4.2 (in particular footnote 2) that Transfer Entropy contains a state-dependent component, due to the synergy between the source $\boldsymbol{Y}_{t-1}^{(\ell)}$ and past history $\boldsymbol{X}_{t-1}^{(k)}$ of the target X. This component is symmetric in the source Y and past of X, however it is viewed as information transfer from Y by TE rather than as storage due to our perspective of information processing here. We can understand this in two ways (see further details in [188]). First, as this perspective focusses on the state transition of X (outlined above), it considers information from the past of X about that transition first (as storage, including any redundant information with the source), and then considers transfer from other sources (which includes that synergistic component with the past of X). Second, the perspective considers transfer as the contribution of the source Y in the *context* of the target past, which as described in Sect. 4.2 has a natural interpretation as conditional MI and naturally includes the synergistic component.

4.2.2.1 Information Transfer and Causality

Returning to Example 4.1 on p. 66, we note that, since the joint process and also Y_t are first-order Markov, while X_t is non-Markov, then following the discussion above we should really consider $\mathbf{T}_{Y\to X}^{(\infty,1)}$ and $\mathbf{T}_{X\to Y}^{(1,1)}$. We have already seen (Eqn. 4.9) that $\mathbf{T}_{X\to Y}^{(1,1)}$ is zero. But in fact $\mathbf{T}_{Y\to X}^{(\infty,1)}$ is also zero! For (if $c \neq 0$) as we take longer and longer histories of X, we gain more and more information about the

[5] We can of course decompose the remaining uncertainty term $\mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}\right)$ in Eqn. 4.17 into further *higher-order* transfer entropy terms, see Sect. 4.2.3.

phase of Y in $\mathbf{A}_X^{(k)}(t)$ (which represents the true *state* of X). In the long history limit, knowing the complete history of X is tantamount to knowing Y, so that as $k \to \infty$, $\mathbf{H}\left(X_t \mid \boldsymbol{X}_{t-1}^{(k)}\right) \to \mathbf{H}(X_t \mid Y_{t-1})$ and $\mathbf{T}_{Y\to X}^{(k,1)} \to 0$ (Fig. 4.1). So, perhaps counter-intuitively, if we take full histories into account, there is no information transfer in *either* direction in Example 4.1. From a computational perspective though, this can be resolved in that the embedded state of X in fact *stores* information about the phase of Y, and thus although Y does *causally* influence X, it does not *transfer* dynamically new information at each update because that causal link serves to maintain *information storage* instead. The concept of causality is typically related to whether *interventions* on a source can be identified to have an effect on the target, rather than whether observation of the source can help *predict* state transitions of the target. The latter concept here is information transfer, whilst the former (causality) may support information transfer or it may support distributed information storage instead. In other words [191]:

Key Idea 19: *Information transfer and causality are related but distinct concepts.*

We refer the reader to [191, 11, 60] for further discussion of the complex relationship between concepts of information transfer and causality. Importantly, the vanishing transfer entropy in the long history limit in this example is due essentially to the deterministic transition of the Y variable. In a more general setting we would not expect $\mathbf{T}_{Y\to X}^{(\infty,1)}$ to vanish if current X has a dependence on the history of Y.

4.2.3 Conditional Transfer Entropy

With many systems there are many interacting variables, so we need to be able to handle additional influences on the pairwise interaction we have discussed so far. When a third (possibly multivariate) process, Z_t, say, is jointly distributed with the processes X_t, Y_t then the *pairwise*, *bivariate* or *apparent* transfer entropy $\mathbf{T}_{Y\to X}$ may report a spurious information flow from Y to X, due to (possibly lagged) joint influences of Z on X and Y (i.e. $Z \to X$ and $Z \to Y$). This is known as a *common driver* effect. Similarly, $\mathbf{T}_{Y\to X}$ may report a spurious information flow from Y to X due to *cascade effects*, e.g. where we actually have $Y \to Z \to X$. Further, $\mathbf{T}_{Y\to X}$ will not detect any synergistic transfer from Y and Z together in these scenarios. It is, however, a simple matter to discount redundant joint influences and include synergies by conditioning on the past of Z. We thus define *conditional transfer entropy* [195, 196, 335]:

Definition 4.4.

$$\begin{aligned}\mathbf{T}_{Y\to X\,|\,Z}^{(k,\ell,m)}(t) &\equiv \mathbf{I}\Big(X_t : \boldsymbol{Y}_{t-1}^{(\ell)} \,\Big|\, \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Z}_{t-1}^{(m)}\Big) \\ &= \mathbf{H}\Big(X_t \,|\, \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Z}_{t-1}^{(m)}\Big) - \mathbf{H}\Big(X_t \,|\, \boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}, \boldsymbol{Z}_{t-1}^{(m)}\Big)\,.\end{aligned} \tag{4.18}$$

Key Idea 20: $\mathbf{T}_{Y\to X\,|\,Z}(t)$ *may be interpreted intuitively as the degree of uncertainty about current X resolved by the past state of Y, X and Z together, over and above the degree of uncertainty about current X already resolved by its own past state* and the past state of *Z*.

We also have:

$\mathbf{T}_{Y\to X\,|\,Z} = 0 \iff$ *X, conditional on its own past* and on the past of *Z, is independent of the past of Y*.

We refer to Z as the *conditioning variable*. Regarding our previous discussion on history lengths, here even if the joint process X_t, Y_t, Z_t is ℓth-order Markov, we would still recommend letting the history length m of the conditioning variable $\to \infty$, since now the joint process X_t, Z_t will generally not be Markov (but we may still use history length ℓ for the source variable Y).

A case of particular practical importance is where we have a system of n jointly distributed processes[6] $\boldsymbol{X}_t = (X_{1,t}, \ldots, X_{n,t})$. Then since, as we have seen, the *pairwise* transfer entropies $\mathbf{T}_{X_j\to X_i}(t)$, $i, j = 1, \ldots, n$ are susceptible to confounds due to common influences of the remaining X_k, an alternative measure of pairwise information flows in the full system $\boldsymbol{X}$ is given by the *pairwise- or bivariate-conditional or complete* transfer entropies [195] (we omit history superscripts, although we should let them all $\to \infty$ here):

Definition 4.5.

$$\begin{aligned}\mathbf{T}_{X_j\to X_i\,|\,\boldsymbol{X}_{[ij]}}(t) &\equiv \mathbf{I}\big(X_{i,t} : X_{j,t-1} \,\big|\, \boldsymbol{X}_{[ij],t-1}\big) \\ &= \mathbf{H}\big(X_{i,t} \,|\, \boldsymbol{X}_{[j],t-1}\big) - \mathbf{H}\big(X_{i,t} \,|\, \boldsymbol{X}_{t-1}\big)\,,\end{aligned} \tag{4.19}$$

where the notation $[\cdots]$ indicates *omission* of the corresponding indices. The conditioning may be limited to only the (other) causal parents of X_i where these are known [195, 191], denoting this variant ${}^{c}\mathbf{T}_{Y\to X}^{(k,l,\mathbf{m})}$. The quantities $\mathbf{T}_{X_j\to X_i\,|\,\boldsymbol{X}_{[ij]}}(t)$, $i \neq j$, may be considered as a directed graph describing the network of information flows between elements of the multivariate system $\boldsymbol{X}$, closely related to the *causal graph* [301, 29] of bivariate-conditional Granger causalities (see below, Sect. 4.4; this is of particular interest in information flow analysis of neural systems—see Sect. 7.3).

[6] Here bold type denotes vector (multivariate) quantities.

Similarly, we may define *collective transfer entropy* [196] as the transfer from some *multivariate* set of n jointly distributed processes $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})$[7] to a specific univariate process, X:

Definition 4.6.

$$\mathbf{T}^{(k,\boldsymbol{\ell})}_{\boldsymbol{Y}\to X}(t) \equiv \mathbf{I}\left(X_t : \boldsymbol{Y}^{(\boldsymbol{\ell})}_{t-1} \,\middle|\, \boldsymbol{X}^{(k)}_{t-1}\right). \tag{4.20}$$

Averaging this over all X in the system under consideration gives the global transfer entropy introduced in Barnett et al. [24] and discussed further in Sect. 5.2.

Finally, with these quantities we may then extend our decomposition of the information content of X_t from Eqn. 4.17, though now considering the information from several sources $\boldsymbol{Y}_t$ (omitting history superscripts on the $Y_{i,t}$ for ease of notation) [196]:

$$\begin{aligned}
\mathbf{H}(X_t) &= \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : X_t\right) + \mathbf{T}^{(k)}_{\boldsymbol{Y}\to X}(t) + \mathbf{H}\left(X_t \mid \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}_{t-1}\right) && (4.21)\\
&= \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : X_t\right) + \left(\sum_i \mathbf{I}\left(X_t : \boldsymbol{Y}_{i,t-1} \mid \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}_{[i\ldots n],t-1}\right)\right)\\
&\quad + \mathbf{H}\left(X_t \mid \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}_{t-1}\right) && (4.22)\\
&= \mathbf{A}^{(k)}_X(t) + \left(\sum_i \mathbf{T}_{Y_i\to X \mid \boldsymbol{Y}_{[i\ldots n]}}(t)\right) + \mathbf{H}\left(X_t \mid \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}_{t-1}\right). && (4.23)
\end{aligned}$$

Each term in the iterative sum over the Y_i above[8] is a transfer entropy term, of increasing order. The sum begins with pairwise TE from Y_1, then adds in conditional TEs from Y_2 through Y_{n-1}, and finally a pairwise- or bivariate-conditional or complete TE from Y_n. Crucially, this equation shows that:

> **Key Idea 21:** *TE terms of various orders are all complementary, and all of these orders of TE terms are required to properly account for the information in the target X_t.*

For example, if we only consider pairwise TE terms, then we would *never* see the synergistic (see Sect. 3.2.3.1) conditional TEs involved in an XOR operation $X_t = Y_{t-1}$ `XOR` Z_{t-1}. Conversely, if we only consider conditional TE terms, then we would *never* see the redundant (see Sect. 3.2.3.1) pairwise TEs involved in a redundant copying operation $X_t = Y_{t-1} = Z_{t-1}$.

[7] We write $\boldsymbol{\ell}$ in vector notation to indicate that potentially different history lengths may be used for each variable in $\boldsymbol{Y}_t$.

[8] Of course, the ordering of the sum over the i is arbitrary, so long as terms already included are conditioned out in later terms.

Key Idea 22: *The term* information dynamics *[195, 196, 198, 182, 199] is used to refer to investigations of the decomposition of information storage and transfer components in Eqn. 4.21–Eqn. 4.23, and also their local dynamics in space and time (see e.g. local transfer entropy in Sect. 4.2.5).*

4.2.4 Source–Target Lag

Transfer entropy may be measured over an arbitrary source–target lag or delay of u time steps [348]:

Definition 4.7.

$$\begin{aligned} \mathbf{T}^{(k,\ell)}_{Y\to X}(t,u) &\equiv \mathbf{I}\Big(X_t : \boldsymbol{Y}^{(\ell)}_{t-u} \,\Big|\, \boldsymbol{X}^{(k)}_{t-1}\Big) \\ &= \mathbf{H}\Big(X_t \mid \boldsymbol{X}^{(k)}_{t-1}\Big) - \mathbf{H}\Big(X_t \mid \boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}^{(\ell)}_{t-u}\Big)\,. \end{aligned} \tag{4.24}$$

Crucially, the lag must only be taken between $\boldsymbol{Y}^{(\ell)}_{t-u}$ and X_t; i.e. $\boldsymbol{X}^{(k)}_{t-1}$ should remain the immediate past of X_t. This is because this form: preserves the computational interpretation of TE as information transfer (see Sect. 4.2.2); is the only relevant option in keeping with Wiener's principle of causality [348]; and crucially, it has been demonstrated that, for a causal relationship $Y \to X$ over a single lag δ, $\mathbf{T}^{(k,\ell)}_{Y\to X}(t,u)$ is maximised at $u = \delta$ [348].

In the previous and the following, we have used $u = 1$ for simplicity, but all formulations can be extended to accommodate an arbitrary delay. Indeed, u should be selected so as to maximise $\mathbf{T}^{(k,\ell)}_{Y\to X}(t,u)$, as described in [348].

4.2.5 Local Transfer Entropy

Since the TE is simply a conditional MI, one can define *local* transfer entropy [195] as a pointwise (local) conditional mutual information (see Eqn. 3.18) from a specific source event state $\boldsymbol{y}^{(\ell)}_{t-1}$ to a specific target event x_t in the context of the specific event state history of the target $\boldsymbol{x}^{(k)}_{t-1}$:

Definition 4.8.

$$\mathbf{t}_{Y\to X}^{(k,\ell)}(t) \equiv \mathbf{i}\left(x_t : \boldsymbol{y}_{t-1}^{(\ell)} \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}\right) \tag{4.25}$$

$$= \log_2 \frac{p\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}\right)}{p\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)},\right)}, \tag{4.26}$$

$$\mathbf{T}_{Y\to X}^{(k,\ell)}(t) = E\left\{\mathbf{t}_{Y\to X}^{(k,\ell)}(t)\right\}. \tag{4.27}$$

$\mathbf{T}_{Y\to X}^{(k,\ell)}(t)$ is the average difference in code length between coding the value x_t given $\boldsymbol{x}_{t-1}^{(k)}$ (under the optimal encoding scheme for X_t given $\boldsymbol{X}_{t-1}^{(k)}$) or coding the value x_t given both $\boldsymbol{x}_{t-1}^{(k)}$ and $\boldsymbol{y}_{t-1}^{(\ell)}$ (under the optimal encoding scheme for X given Y and Z), while $\mathbf{t}_{Y\to X}^{(k,\ell)}(t)$ represents this difference in such code lengths for any specific events $\{x_t, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}\}$ under these schemes. As such:

Key Idea 23: *The local transfer entropy tells us about the* dynamics *of information transfer in time.*

We will see specific examples of such dynamics in Chap. 5.

The local transfer entropy may be either positive *or* *negative* (with the source $\boldsymbol{y}_{t-1}^{(\ell)}$ being either informative or *misinformative* respectively) for a specific event set $\{x_t, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}\}$, as explained in Sect. 3.2.2 and Sect. 3.2.3 for local MI and local conditional MI values. Further examples are given in Sect. 5.1.

Of course, the conditional TE (Eqn. 4.18), complete TE (Eqn. 4.19) and collective TE can all be localised in a similar manner using the local conditional MI in Eqn. 3.18 [195].

4.3 Transfer Entropy Estimators

The same issues which plague the estimation of entropy and mutual information discussed in Chap. 3 plague transfer entropy to an even greater extent due to its generally larger dimensionality. To some degree, estimators for MI and conditional MI introduced in Sect. 3.4 may be directly applied to estimate the transfer entropy. One should keep in mind though that, as we saw in Sect. 3.4, the straightforward plug-in entropy estimator, in which we estimate the probabilities from the counts and apply Eqn. 3.2, has a positive bias and behaves less well than other, indirect estimators. In this section then, we describe direct estimation of the transfer entropy.

Finding good estimators is an open research area, and the reader is recommended to use some of the available toolboxes described in Sect. 4.3.3 at the outset of a project.

4.3.1 KSG Estimation for Transfer Entropy

Considerable work has gone into the direct estimation of mutual information, using kernels (Sect. 3.4.1.4) and the Kozachenko–Leonenkov entropy estimator, in the KSG (Kraskov) estimator (Sect. 3.4.2.2). As a result it might be tempting to use the mutual information to estimate the transfer entropy. Indeed, Kraskov [167] initially suggested that TE could be computed as the difference between two mutual information terms:

$$\mathbf{T}^{(k,\ell)}_{Y\to X}(t) = \mathbf{I}\left(X_t, \boldsymbol{X}^{(k)}_{t-1} : \boldsymbol{Y}^{(\ell)}_{t-1}\right) - \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : \boldsymbol{Y}^{(\ell)}_{t-1}\right) \tag{4.28}$$

$$= \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1}, \boldsymbol{Y}^{(\ell)}_{t-1} : X_t\right) - \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : X_t\right), \tag{4.29}$$

and similarly it is easy to verify that

$$\mathbf{T}^{(k,\ell)}_{Y\to X}(t) = \mathbf{I}(\boldsymbol{X}^{(k)}_{t-1} : \boldsymbol{Y}^{(\ell)}_{t-1} : X_t) - \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : \boldsymbol{Y}^{(\ell)}_{t-1}\right) - \mathbf{I}\left(\boldsymbol{X}^{(k)}_{t-1} : X_t\right). \tag{4.30}$$

The above expressions are *exact*, theoretically, but numerically there can be problems. So if we calculate the mutual information using the KSG estimator (Sect. 3.4.2.2), then there is a positive bias, causing the TE to be over-estimated. Using a difference of mutual information terms here leads to an over-estimate, because the nearest-neighbour distances would be calculated separately for each term. Thus the balls are smaller for the two-dimensional MIs, leading to a smaller estimate for the effective nearest-neighbour count.

This issue has been addressed by extending the KSG algorithm for direct estimation of the conditional mutual information [93, 110, 337, 350], as alluded to in Sect. 3.4.2.3. To understand this, we demonstrate the application of the Kraskov approach (algorithm 1) directly for $\mathbf{I}(X : Y \mid Z)$. In terms of entropies we have

$$\mathbf{I}(X : Y \mid Z) = \mathbf{H}(X,Z) + \mathbf{H}(Y,Z) - \mathbf{H}(X,Y,Z) - \mathbf{H}(Z). \tag{4.31}$$

Applying the same logic as Kraskov et al. did for mutual information, but here with a single ball for the Kth nearest neighbour in the *joint distribution*, $\{x,y,z\}$, leads to (for algorithm 1)

$$\mathbf{I}^{(1)}(X : Y \mid Z) = \psi(K) - E\left\{\psi(n_{xz}) - \psi(n_{yz}) + \psi(n_z)\right\}. \tag{4.32}$$

in *nats*. Here ε is the (max) norm to the Kth nearest neighbour in the joint space $\{x,y,z\}$ for each given test point, n_z is the neighbour count strictly within norm ε in the z marginal space, and n_{xz} and n_{yz} are the neighbour counts strictly within (max) norms of ε in the joint $\{x,z\}$ and $\{y,z\}$ spaces, respectively.

Similarly, following KSG algorithm 2, $\{\varepsilon_x, \varepsilon_y, \varepsilon_z\}$ are set separately to the marginal distances to the Kth nearest neighbour in the joint space $\{x,y,z\}$ for each given test point, and one then counts $\{n_z, n_{xz}, n_{yz}\}$ within or on these widths to obtain [350]

$$\mathbf{I}^{(2)}(X:Y\mid Z)=\psi(K)-\frac{2}{K}+E\left\{\psi(n_z)-\psi(n_{xz})+\frac{1}{n_{xz}}-\psi(n_{yz})+\frac{1}{n_{yz}}\right\} \quad (4.33)$$

in *nats*.

As a conditional MI (cf. Eqn. 4.11), direct estimation of transfer entropy may *then* be performed via these algorithms [110, 337, 350]. Crucially, the search for nearest neighbours may be performed using optimised algorithms in $\mathrm{O}(KN\log N)$ time (for N samples) instead of $\mathrm{O}\left(KN^2\right)$ for a naive all-to-all neighbour search over N samples (see [183]).

Open Research Question 3: *What are the best estimators for different probability distributions and for large dimensionality?*

4.3.1.1 Non-stationarity

When the statistics are non-stationary, the formulae for TE still apply, taken over ensembles. In some situations, one has access to or is able to generate such an ensemble, e.g. see TE analysis of ensembles of repeated trials of event-driven stimulus in neuroscientific experiments in [110, 350, 180, 363]. In other practical situations such ensembles would often not be available. For example, in financial time series, there is only one time record of the price of shares and the share index for a given stock exchange. Thus the only practical course of action is to use time windows of a small enough size that the statistics are (approximately) stationary over the window. But a small window may make the estimation with such a small number of data points very unreliable.

Open Research Question 4: *Are there better methods for calculating TE, suitable for real data, for non-stationary systems without ensemble data?*

4.3.2 *Symbolic Transfer Entropy*

One way around handling continuous distributions with relatively small number of data points is a different form of discretisation or binning, *symbolic transfer entropy*, introduced by Staniek and Lehnertz [313]. The idea here is to take the embedding dimension $\mathbf{m}$ (see Sect. 2.3.5; i.e. k for $\boldsymbol{X}^{(k)}$) for the time series in question and for each data value look at the *ordering* of the current and previous $\mathbf{m}-1$ values and assign a symbol according to which permutation of magnitudes it corresponds.

Thus for three values, $x_3 > x_2 > x_1$ would have a different symbol to $x_2 > x_3 > x_1$. The statistics of occurrence of the symbols are then combined and these probability distributions used to calculate the entropy of the series, with entropy combinations used for MI and TE, etc.

This is a particularly fast approach, since it effectively computes a discrete entropy after the ordinal symbolisation. It is important to note, however, that it is model based, assuming that all relevant information is in the ordinal relationship between the variables. This is not necessarily the case in the variables we are analysing, and can lead to misleading results, as has been demonstrated by Wibral et al. [348].

4.3.3 Open-Source Transfer Entropy Software

A number of existing open-source toolkits are available for computing the transfer entropy empirically from time-series data, as described in the following. For each toolkit, we describe its purpose, the type of data it handles, and which estimators are implemented. At the risk of including bias, the first two toolkits presented are associated with authors of this book.

The MVGC (multivariate Granger causality toolbox)[9] (GPL v3 licence) by Barnett (an author of this book) and Seth [26] provides general-purpose calculation of the Granger causality for MATLAB (MVGC also requires the MATLAB Statistics, Signal Processing Toolbox). MVGC allows specification of embedding dimension, but not source–target delay parameters.

The Java Information Dynamics Toolkit (JIDT)[10] (GPL v3 licence) by Lizier (an author of this book) [183] provides general-purpose calculation of the transfer entropy on a variety of platforms (while written in Java, it is usable in MATLAB, Octave, Python, R etc.). JIDT implements TE and conditional TE, plus a range of related measures (entropy, MI, conditional MI, AIS and more). This is done using a variety of estimator types (discrete/binned, Gaussian, box-kernel and KSG including fast nearest-neighbour search and parallel computation). JIDT allows specification and auto selection of embedding dimension and source–target delay parameters, and adds capabilities to compute local information-theoretic values (e.g. local transfer entropy, see Sect. 4.2.5), collective TE and statistical significance testing (see Sect. 4.5.1). Several demonstrations of computing TE using JIDT are distributed with the toolkit, and some are described here in Chaps. 5 and 7.

TRENTOOL[11] (GPL v3 licence) by Lindner et al. [180] is a MATLAB toolbox designed from the ground up for transfer entropy analysis of (continuous) neural data, utilising the FieldTrip [250] data format for electroencephalography (EEG), magnetoencephalography (MEG), and local field potential (LFP) recordings. In particular, it is designed for performing effective network or connectivity analysis (see

[9] http://www.sussex.ac.uk/sackler/mvgc/

[10] http://jlizier.github.io/jidt/

[11] http://www.trentool.de

Sect. 7.2) between the input variables, including statistical significance testing of TE results (see Sect. 4.5.1) and other steps to deal with volume conduction and identify cascade or common driver effects in the inferred network. TRENTOOL automates selection of embedding parameters for input time-series data and for source–target lags, and implements KSG estimation via fast nearest-neighbour search, parallel computation and graphics processing unit (GPU)-based algorithms [363].

The MuTE toolbox[12] by Montalto et al. [230] (CC-BY license) implements TE estimation for MATLAB. MuTE is capable of computing conditional TE and includes a number of estimator types (discrete/binned, Gaussian, and KSG including fast nearest-neighbour search). It also adds non-uniform embedding (see Faes et al. [85]), methods to assist with embedding parameter selection, and statistical significance testing.

TIM[13] (GNU Lesser GPL licence) by Rutanen [292] provides C++ code (callable from MATLAB) for general-purpose calculation of a wide range of information-theoretic measures on continuous-valued data. TIM implements entropy (Shannon, Rényi and Tsallis variants), Kullback–Leibler divergence, MI, conditional MI, TE and conditional TE. TIM includes various estimators for these, notably with KSG estimators (using fast nearest-neighbour search). Estimators are also included for multi-dimensional variables.

The Transfer Entropy Toolbox (TET)[14] (BSD licence) by Ito et al. [142] provides TE analysis of spiking (binary, discrete) data for MATLAB. TET allows specification of embedding dimension and source–target delay parameters.

Users should make a careful choice of which toolkit suits their requirements, considering data types, estimators and application domain. For example, TRENTOOL is dedicated to effective network inference in neural imaging data, and so is an ideal tool for that application. For more general-purpose applications, a toolkit such as MVGC or JIDT would be more suitable.

4.4 Relationship with Wiener–Granger Causality

As mentioned in the introduction to this chapter, transfer entropy is closely related to and (arguably) shares a common history with Wiener–Granger causality (Granger causality for short) [354, 114, 112, 105, 285]. It was not, however, till [22, 23] that the precise relationship between the concepts was formally elucidated. In this section we provide a brief introduction to the conceptual, operational and inferential basis of Granger causality. We then examine in more detail its relationship with transfer entropy.

[12] http://figshare.com/articles/MuTE_toolbox_to_evaluate_Multivariate_Transfer_Entropy/1005245/1

[13] http://www.cs.tut.fi/%7etimhome/tim/tim.htm

[14] http://code.google.com/p/transfer-entropy-toolbox/

4.4.1 Granger Causality Captures Causality as Predictive of Effect

Firstly, however, no mention of Granger causality can avoid some remarks as to the notion of *causality* intended by the nomenclature. Causality in the Wiener–Granger sense is perhaps best summarised as [114]

Key Idea 24: *Granger causality is based on the premise that cause precedes effect, and a cause contains information about the effect that is unique, and is in no other variable.*

It would seem to be the case that, to many people, this notion of causality fails to tally with preconceived ideas based on distinctly different premises (in particular *interventionist* approaches [261, 11, 191, 60]; see Sect. 4.2.2.1). We do not intend to engage in this debate here, which we feel has generated rather more heat than light. Rather, we are happy instead to accept Granger causality at face value as *a* (as opposed to *the*) notion of causality—in particular, of predictive effect—and allow Granger himself the last (somewhat jaundiced) word on the matter:

> At that time, I had little idea that so many people had very fixed ideas about causation, but they did agree that my definition was not *true causation* in their eyes, it was only *Granger causation*. I would ask for a definition of true causation, but no one would reply. However, my definition was pragmatic and any applied researcher with two or more time series could apply it, so I got plenty of citations. Of course, many ridiculous papers appeared.
>
> Clive W. J. Granger, Nobel Lecture, December 8, 2003 [114]

4.4.2 Definition of Granger Causality

For simplicity we consider just the bivariate case of two jointly stationary, possibly multivariate, stochastic processes X_t, Y_t—as for transfer entropy, Granger causality extends (in a reasonably straightforward manner) to the non-stationary/conditional cases. In its *purest* (though not historically original) form, the essence of the idea is surprisingly close to that of transfer entropy. Let $F\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}\right)$ denote the distribution function of the target variable X conditional on the joint (k,ℓ)-history $\boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}$ of both itself and the source variable Y, and let $F\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}\right)$ denote the distribution function of X_t conditional on just its own k-history. Then [112, 115] variable Y is said to Granger-cause variable X (with lags k,ℓ) iff

$$F\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}\right) \neq F\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}\right). \tag{4.34}$$

In other words:

Key Idea 25: *Y Granger-causes X iff X, conditional on its own history, is* not *independent of the history of Y.*

The connection with transfer entropy is clear: in fact (4.34) holds precisely when $\mathbf{T}^{(k,\ell)}_{Y\to X} \neq 0$. Thus Transfer Entropy might be construed as a non-parametric test statistic for *pure* Granger causality! But this is not the historical path that the development of Granger causality took. In [113] Granger remarks regarding (4.34) *The general definition [...] is not operational, in that it cannot be used with actual data. To become operational, a number of constraints need to be introduced.* In fact Granger had already—apparently inspired by an idea due to Wiener [354]—operationalised the concept via *parametric predictive modelling*, and the non-parametric, information-theoretic version was (rather surprisingly, one might think) to wait another 40 years to emerge in coherent form.

Granger's parametric formulation was, specifically, based on linear vector autoregressive (VAR) modelling [126, 207]. X_t, Y_t are assumed to be multivariate real-valued, zero-mean, jointly stationary stochastic processes, subject to some restrictions, which we clarify below. We are then asked to consider[15] the nested VAR models

$$X_t = A_1 \cdot X_{t-1} + \ldots + A_k \cdot X_{t-k} + B_1 \cdot Y_{t-1} + \ldots + B_\ell \cdot Y_{t-\ell} + \varepsilon_t, \tag{4.35}$$

$$X_t = A'_1 \cdot X_{t-1} + \ldots + A'_k \cdot X_{t-k} + \varepsilon'_t. \tag{4.36}$$

The parameters of the models are the VAR coefficient matrices A_i, B_j, A'_i and the covariance matrices $\Sigma \equiv \mathbf{c}(\varepsilon_t), \Sigma' \equiv \mathbf{c}(\varepsilon'_t)$ where $\varepsilon_t, \varepsilon'_t$ are the *residuals*, assumed to be serially (though not necessarily contemporaneously) uncorrelated; (4.35) and (4.36) are referred to, respectively, as the *full* and *reduced* models. There are now two approaches, which turn out to be roughly equivalent.

The first—Granger's original approach (via Wiener)—views (4.35), (4.36) as *predictive* models for the target variable X in terms of, respectively, the joint past of itself and the source variable Y (full model), and its own past only (reduced model). Then the $Y \to X$ Granger causality statistic stands to quantify the degree to which the full model yields a *better* prediction of the target variable (perhaps in the least-squares sense) than the reduced model. Standard linear prediction theory [126, 207] suggests that this should be measured by some R^2-like statistic based on the ratio of residuals variances. Following Geweke [105], the most convenient form for the Granger causality statistic (for reasons which will become clear) is given by

Definition 4.9.

$$F^{(k,\ell)}_{Y\to X} \equiv \log \frac{|\Sigma'|}{|\Sigma|}, \tag{4.37}$$

where $|\cdot|$ denotes the matrix determinant.

[15] Our presentation here is closer to that of Geweke [105], who developed the now-standard modern approach to Granger-causal inference, in both the time and (see below) spectral domains.

(The determinant of a residuals covariance matrix is sometimes known as the *generalised* variance, as opposed to the *total* variance, i.e. sum of variances.) In Definition 4.9 the model parameters are assumed to have been chosen [e.g. by ordinary least-squares (OLS)] to minimise the total variance (or, equivalently, as it turns out, the generalised variance) of the respective models.[16]

The second, perhaps more principled, approach, is within a *maximum-likelihood* (ML) framework [82]. Here we note that $F_{Y\to X}$ (again we drop the superscripts if convenient) is precisely the *log-likelihood ratio* statistic for the model (4.35) under the null hypothesis

$$H_0 : B_1 = B_2 = \ldots = B_\ell = 0. \tag{4.38}$$

Note that, given that X_t, Y_t is described by the model (4.35), the null hypothesis (4.38) is precisely the negation of condition (4.34) for *non*-causality. An immediate payoff of the ML approach is that we have an (asymptotic) expression for the sample distribution of the statistic $F_{Y\to X}$ as a χ^2 with degrees of freedom equal to the difference in number of free parameters between the full and reduced models.[17]

A further property of Granger causality is that (unlike transfer entropy) it extends naturally to the spectral domain [105, 106], so that causal interactions may be decomposed by frequency.

In [25] it is also shown that the Granger causality statistic (in both time and frequency domains) is on the analytical level invariant under arbitrary stable invertible filtering. However, it is also demonstrated that, for empirical estimation from time-series data, (invertible) filtering will, in general, degrade Granger-causal inference. The reason for this is that filtering a VAR process will generally increase the VAR model order and/or induce a moving average (MA) component, resulting in poor VAR modelling and an increased number of model parameters. This is a serious practical issue, particularly in applications of Granger causality to neurophysiological data (Sect. 7.3), where time series are routinely filtered as a pre-processing step, often with the intention of eliminating frequency bands deemed biophysically implausible, or for suppression of artefacts. It is also not uncommon in the neuroscience literature to find that data has been band-filtered with the stated objective of estimating Granger causality restricted to a specific frequency band. [25] show that such pre-filtering not only fails to achieve this goal, but may well increase the incidence of false positives and false negatives in causal inference. Rather, *band-limited* Granger causality should be calculated by integrating frequency-domain Granger causality over the requisite frequncy range. [25] recommend that pre-filtering be kept to an absolute minimum required, e.g., to achieve better stationarity; thus notch filtering to suppress line noise, or high-pass filtering to eliminate slow transients, is acceptable if the alternative is failure of VAR modelling due to non-stationarity.

---

[16] We note that Granger himself considered the total rather than generalised variance for his test statistic. For further discussion on the preferability of the generalised variance, see [29].

[17] If the target X is *univariate*, the sample distribution of the R^2 statistic $\exp(F_{Y\to X}) - 1$ is asymptotically described by an F-distribution, which has somewhat fatter tails than the corresponding χ^2 and, in this case, yields better statistical inference. This is, presumably, the origin of the conventional F notation for the Granger statistic.

Open Research Question 5: *Is transfer entropy invariant under arbitrary* non-linear *invertible causal filtering?*

It seems likely that this corresponding result for transfer entropy ought to be obtained, although further technical conditions may be required.

4.4.3 Maximum-Likelihood Estimation of Granger Causality

How should the statistic $F_{Y\to X}$ be applied for time-series data? Standard VAR model fitting techniques (such as OLS or Levinson–Wiggins–Robinson (LWR) algorithms [178, 355, 232]) may be deployed to derive least-squares/ML estimates for VAR parameters of the full and reduced regressions, in particular the covariance matrices Σ, Σ'. Firstly—as for transfer entropy—we will need to select suitable numbers of historical lags (k,ℓ)—the *model orders*, in the VAR framework—for the regressions.[18] Again, the ML framework is useful here since the generalised residuals variance is also the likelihood for a ML estimate of the corresponding regression, and may be supplied to popular model order estimation criteria such as the Akaike or Bayesian information criteria [221]. The covariance matrices Σ, Σ' for the optimal model order may then be plugged directly into Eqn. 4.37. If the amount of data is sufficient, the appropriate theoretical asymptotic χ^2 (or F) distribution may be used for statistical inference (for short time series or high model orders, standard sub-sampling or surrogate data techniques may be more reliable).

Barnett et al. [22] prove the following theorem:

Theorem 4.1. *If the joint process X_t, Y_t is* Gaussian *(more precisely, if any finite subset $\{X_{t_1}, Y_{t_2} : (t_1, t_2) \in S\}$ of the variables is distributed as a multivariate Gaussian) then there is an exact equivalence between the Granger causality and transfer entropy statistics:*

$$\mathbf{T}^{(k,\ell)}_{Y\to X} = \tfrac{1}{2} F^{(k,\ell)}_{Y\to X}. \tag{4.39}$$

The proof is rather straightforward, and is based on the facts that: (i) given an arbitrary vector linear regression $U = A \cdot V + \varepsilon$, the least-squares/ML estimate for the residuals covariance matrix $\mathbf{c}(\varepsilon)$ is given by the *partial covariance*

$$\mathbf{c}(U \mid V) \equiv \mathbf{c}(U) - \mathbf{c}(U,V)\,\mathbf{c}(V)^{-1}\,\mathbf{c}(V,U), \tag{4.40}$$

and (ii) the conditional entropy of jointly multivariate Gaussian variables U, V is

$$\mathbf{H}(U \mid V) = \tfrac{1}{2}\log(|\mathbf{c}(U \mid V)|) + \tfrac{1}{2} n \log(2\pi e), \tag{4.41}$$

[18] On a technical point, we note that the same target model order k should be used in both full and reduced regressions and should, preferably, be estimated from the *reduced* regression. For the reasons, see [26].

where $n = \dim(U)$. Then taking $U = X_t$ and $V = \left(\boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{Y}_{t-1}^{(\ell)}\right)$ (full regression) and $V = \boldsymbol{X}_{t-1}^{(k)}$ (reduced regression), respectively, the result follows directly from Definition 4.2 for transfer entropy and Definition 4.9 for Granger causality. This result was subsequently extended (for VAR models) to various generalised Gaussian/exponential distributions [138] and finally by Barnett et al. [23] to a very general class of predictive models in a ML framework (see also [285]). The chief result in [23] may be stated as:

Theorem 4.2. *Suppose that the conditional distribution function of the target variable X_t on its own* entire *past and that of the source variable Y_t satisfies the order-(k,ℓ) partial Markov model*

$$F\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(\infty)}, \boldsymbol{y}_{t-1}^{(\infty)}\right) = f\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}; \boldsymbol{\theta}\right), \tag{4.42}$$

where $\boldsymbol{\theta}$ is a (finite-dimensional) parameter vector. Then, under assumption that the model (4.42) *is identifiable and well-specified, and that a certain (non-restrictive) ergodicity condition is satisfied, the* ML transfer entropy estimator

$$\widehat{\mathbf{T}}_{Y\to X}^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right) \equiv -\frac{1}{N-k}\log \Lambda^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right) \tag{4.43}$$

converges almost surely to the actual transfer entropy:

$$\widehat{\mathbf{T}}_{Y\to X}^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right) \xrightarrow{a.s.} \mathbf{T}_{Y\to X}^{(k,\ell)} \tag{4.44}$$

as the sample size $N \to \infty$, where $\Lambda^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right)$ is the likelihood ratio *for the model* (4.42) *and the nested model defined by the null hypothesis [cf.* (4.38)*]*

$$H_0 : f\left(x_t \,\middle|\, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_{t-1}^{(\ell)}; \boldsymbol{\theta}\right) \text{ does not depend on } \boldsymbol{y}_{t-1}^{(\ell)}. \tag{4.45}$$

Theorem 4.2 states, in other words, that the ML estimator $\widehat{\mathbf{T}}_{Y\to X}^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right)$ of (4.43) is a *consistent estimator* for the actual transfer entropy $\mathbf{T}_{Y\to X}^{(k,\ell)}$. As a corollary, the scaled estimator $2(N-k)\widehat{\mathbf{T}}_{Y\to X}^{(k,\ell)}\left(\boldsymbol{x}_t^{(N)}, \boldsymbol{y}_t^{(N)}\right)$ has an asymptotic $\chi^2(d)$ distribution under the null hypothesis H_0 (zero transfer entropy), where the number of degrees of freedom d is the difference between the number of free parameters in the unrestricted and null models, while under the alternative hypothesis (non-zero transfer entropy) the asymptotic distribution is non-central $\chi^2(d;\lambda)$ with non-centrality parameter $\lambda = 2(N-k)\mathbf{T}_{Y\to X}^{(k,\ell)}$. For a linear finite-order VAR model, we recover the result of Theorem 4.1, albeit only asymptotically.

Key Idea 26: *Theorem 4.2 blurs the boundaries between Granger causality and transfer entropy; thus we might consider the ML estimator* (4.43) *as defin-*

ing a generalised (non-linear) Granger causality or, alternatively, a parametric transfer entropy statistic.

The theorem has far-reaching consequences: if we may assume (perhaps on domain-specific or empirical grounds) that a predictive model of the form (4.42) is appropriate to our data and if, in addition, efficient algorithms are available for ML parameter estimation, then the ML estimator of Theorem 4.2 may well prove easier to calculate and more efficient than direct entropy/mutual information-based estimators (cf. Sect. 4.3). Furthermore, a χ^2 sampling distribution becomes available for free. This suggests potential principled extensions of Granger causality beyond simple linear VAR modelling to a range of standard, well-understood, parametric predictive stochastic models, such as VARMA (vector autoregressive moving-average), VARFIMA (vector autoregressive fractionally-integrated moving-average) and various flavours of GARCH (generalised autoregressive heteroscedastic) models. A particular case of interest is finite-state discrete Markov chain models; here, considering the Markov transition probabilities *themselves* as model parameters, the ML parameter estimators are just the standard plug-in estimators for these probabilities, and the *naïve* plug-in transfer entropy estimator (Sect. 3.4.1.1) is seen to have a χ^2 distribution [23]. Further discussion on this is provided in Sect. 4.5.1.

Open Research Question 6: *Can more sophisticated estimators (kernel-based, adaptive partitioning, k-nearest neighbour, etc., see Sect. 3.4.2) be expressed as predictive parametric models, to which Theorem 4.2 applies?*

4.4.4 Granger Causality Versus Transfer Entropy

It should be clear by now that Granger causality (or perhaps more broadly the generalised Granger causality of Theorem 4.2) offers some obvious advantages over non-parametric transfer entropy as a data-driven, time-directed, functional analysis technique; in particular the ease and efficiency of VAR model parameter estimation as compared with the difficulties (and comparative statistical inefficiency) of entropy/mutual information estimation, as well as the existence of known theoretical sampling distributions for statistical inference. Coupled with the equivalence with transfer entropy for Gaussian processes, why, then, should we bother with (non-parametric) transfer entropy at all? The answer depends largely on the nature of the data and the stochastic generative processes underlying it. Obviously some classes of data (e.g. discrete data with low-cardinality state spaces) are inherently unsuited to VAR modelling.

Other reasons relate to two common misconceptions regarding Granger causality. The first is that Granger causality *can only detect linear dependencies* between variables. That this is by no means the case stems from a "universality" of VAR models, in the following sense: by the celebrated Wold decomposition theorem [80, 128], a broad class of (covariance stationary) stochastic processes—*including many processes with non-linear feedback between variables*—have a moving average (MA) representation. If this representation is, furthermore, *square-summable* and *invertible*, then the process also admits an (albeit, in general infinite-dimensional) VAR representation. Under some further spectral conditions (which ensure that subprocesses are also representable as VARs) the process will then be amenable to Granger causality analysis—see [290] and [105] [in particular eq. (2.4)] for technical details. We note that the invertibility condition precludes, for instance, stationary invertible processes that have been filtered by non-invertible linear filters (e.g. finite differencing[19]). In these cases it is possible that transfer entropy may still yield meaningful results, although little appears to be known on this issue.

The second misconception is that Granger-causal inference is viable only for *Gaussian* processes. Of course we should, at the risk of model mis-specification, be cautious that our VAR model-fitting techniques do not depend too heavily on Gaussian assumptions. As to statistical inference, the standard large-scale theory for ML estimation [242, 243, 356, 341] holds for non-Gaussian processes, although asymptotic convergence of ML estimators to the appropriate χ^2 may suffer.

Perhaps more pertinently, though, even if the data satisfy the technical conditions for a linear VAR model amenable to Granger-causal analysis, it does not follow that the VAR model will necessarily be *parsimonious*. In practice, especially with limited data, this may manifest itself in unacceptably high empirical model orders and poor model fit, which are likely to compromise statistical inference. This may be the case, for instance, for highly non-linear and/or non-Gaussian data, for data with a strong moving average component, or for data which is fractionally integrated [13] or highly heteroscedastic [126]. For such data, in lieu of an appropriate and tractable parametric model (in the sense of Theorem 4.2), non-parametric transfer entropy may well be preferable—for further discussion on this issue see [23].

Key Idea 27: *Finally, we should stress that, for non-Gaussian processes, transfer entropy and Granger causality are simply* not measuring the same thing!

As such, if the intention is explicitly to measure information flow—as opposed to causality in the Granger–Wiener sense—we must use transfer entropy.

[19] Finite differencing is sometimes used to improve stationarity of time series, but in fact renders the resulting process inappropriate for direct Granger-causal analysis. A non-stationary process for which the (perhaps multiply) finite-differenced process is stationary is known as a *unit root* process. Granger causality may, in fact, be estimated for such processes via *co-integration* models, such as vector-error correction (VECM) models. We refer the reader to [207] for the theoretical background.

4.5 Comparing Transfer Entropy Values

A question which naturally arises is whether measurements of transfer entropy in two different systems are directly comparable or not. In particular—given that TE measurements contain bias—is any one TE measurement statistically different from zero or not? Also, different systems may have very different types of dynamics—should we normalise the TE measurements somehow before comparing them? We consider these types of questions in the following.

4.5.1 Statistical Significance

In *theory*, the TE between two variables Y and X with no directed relationship (conditional on the past of X) is equal to 0. In *practice*, where the TE is empirically measured from a finite number of samples N, a bias of a non-zero measurement may result even where there is no such (directed) relationship. Even for bias-corrected estimators, statistical fluctuations give rise to a variance in our measurement here. So a key question is whether a given empirical measurement of TE is statistically different from 0, and implies a directed relationship.

To address this, standard sub-sampling techniques such as permutation testing and bootstrapping may be employed for significance testing and estimation of confidence intervals for the transfer entropy [56, 335, 337, 180, 187, 23, 350, 183]. This is done by forming a *null hypothesis* H_0 that there is no such relationship, and making a test of evidence (our original measurement) in support of that hypothesis. To perform such a test, we need to know what the *distribution* for our measurement would look like if H_0 was true, and then evaluate a p-value for sampling our actual measurement from this distribution. If the test fails, we may accept the alternate hypothesis that there is a (directed) relationship.

For a TE measurement $\widehat{\mathbf{T}}^{(k,\ell)}_{Y\to X}$, we consider the distribution of *surrogate* measurements $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ under the assumption of H_0. Here, Y^s represents *surrogate* variables for Y generated under H_0, which have the same statistical properties as Y, but any potential (conditional) directed relationship with X is destroyed. Specifically, this means that $p(x_t \mid \boldsymbol{x}^{(k)}_{t-1}, \boldsymbol{y}^{(\ell)}_{t-1})$ in Eqn. 4.11 is empirically distributed as $p(x_t \mid \boldsymbol{x}^{(k)}_{t-1})$ (with $p(\boldsymbol{y}^{(\ell)}_{t-1})$ retained).

In some situations, we can compute the surrogate distribution $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ analytically. As described in Sect. 4.4, for Gaussian estimation the null $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ (in *nats*) is asymptotically $\chi^2/2N$ distributed with $\ell d_X d_Y$ degrees of freedom (for dimensionalities d_X and d_Y of potentially multivariate X and Y) [105, 23]. Similarly, for discrete X and Y with cardinality M_X and M_Y, $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ (in *bits*) is asymptotically $\chi^2/(2N\log 2)$ distributed with $(M_X-1)(M_Y^\ell-1)M_X^k$ degrees of freedom [23] (building on [47, 58]).

We must emphasise that such analytic distributions are *asymptotically* correct as the number of samples $N\to\infty$, and the approach is slower for increasing dimen-

sionality of the variables or for discrete variables with skewed distributions (e.g. see [183]). For use in statistical significance testing, the role of a given finite N in the form of the distribution is a crucial factor, so if using an analytic distribution for $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$, then one needs to be careful that it is not too divergent from the true underlying distribution for the given N. Further, analytic surrogate distributions for other estimators remain an open topic of research (see Open Research Question 26).

As such, the distribution of $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ in these cases is empirically computed by sub-sampling techniques such as permutation testing or bootstrapping [56, 335, 337, 180, 187, 350], i.e. manually creating a large number of surrogate time-series pairs $\{Y^s, X\}$ (which meet the statistical form described above), and computing a population of $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ values. Directly shuffling the time series Y to create the set of Y^s is *not* valid, since it destroys the $\boldsymbol{y}^{(\ell)}_{t-1}$ samples (unless $\ell = 1$). It is valid however to: shuffle (or redraw) the $\boldsymbol{y}^{(\ell)}_{t-1}$ amongst the set of $\{x_t, \boldsymbol{x}^{(k)}_{t-1}, \boldsymbol{y}^{(\ell)}_{t-1}\}$ tuples; rotate the Y time series (where we have stationarity); or swap sample source time series Y_i between different trials i in an ensemble approach [337, 350, 180, 363].[20]

Finally, with the distribution of $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ determined, one can compute a p-value for sampling the measured $\widehat{\mathbf{T}}^{(k,\ell)}_{Y\to X}$ under H_0 and compare it with some threshold α.

We will discuss in Sect. 7.2 the important application of such tests of statistical significance in *effective network inference* from multivariate time-series data.

4.5.2 Normalising Transfer Entropy

One often wishes to compare TE values between different pairs of variables—e.g. between which pair of brain regions is most information transferred in a given functional magnetic resonance imaging (fMRI) brain image recording? Yet different systems—or even different pairs of variables in the same system—experience different types of dynamics, and perhaps one should correct somehow for these differences before making comparisons. Here we consider a number of suggestions on how to make such corrections, or *normalise*, TE values.

One key method here is bias correction, since bias could be higher or lower under different dynamics. While some estimators include such correction automatically (e.g. the KSG estimator, see Sect. 4.3.1), this may be performed for other estimators by computing the null distribution $\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}$ as per Sect. 4.5.1 and then subtracting out the mean $E\left\{\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}\right\}$ of this distribution. Marschinski and Kantz [216] introduce this as the *effective transfer entropy*.

Another step is to consider TE as a *fraction* of the maximum value that it could potentially take under the given dynamics. At first glance, one may consider this maximum to be the entropy in the next value of the target; however it is actually

[20] Extension to conditional TE is straightforward by considering the conditioned variable jointly with the past target state $\boldsymbol{x}^{(k)}_{t-1}$.

capped by the entropy rate of the target, $\mathbf{H}'_X(t)$ (Sect. 4.2.2), as one may ascertain from Eqn. 4.11. As such, Gourévitch and Eggermont [111] proposed the *normalised transfer entropy* as:

$$ {}^{n}\mathbf{T}^{(k,\ell)}_{Y\to X} = \frac{\widehat{\mathbf{T}}^{(k,\ell)}_{Y\to X} - E\left\{\widehat{\mathbf{T}}^{(k,\ell)}_{Y^s\to X}\right\}}{\mathbf{H}'_X(t)}, \tag{4.46} $$

which first removes the bias (as per the effective TE, above) and then normalises by the entropy rate $\mathbf{H}'_X(t)$. Gourévitch and Eggermont explain that this represents the fraction of information in the target X not explained by its own past that is explained by Y in conjunction with that past. This normalisation has been used for example in various studies in computational neuroscience [244, 323].

4.6 Information Transfer Density and Phase Transitions

To gauge the *density* of information flows within a system $\boldsymbol{X}$, one can simply use the *average pairwise transfer entropy*:

$$ \mathbf{T}_{pw}(\boldsymbol{X}) \equiv \frac{1}{n(n-1)} \sum_{i\neq j} \mathbf{T}_{X_j\to X_i}(t), \tag{4.47} $$

or the *average bivariate-conditional transfer entropy*:

$$ \mathbf{T}_{bv}(\boldsymbol{X}) \equiv \frac{1}{n(n-1)} \sum_{i\neq j} \mathbf{T}_{X_j\to X_i\,|\,\boldsymbol{X}_{[ij]}}(t). \tag{4.48} $$

Note that, since transfer entropies are non-negative, $\mathbf{T}_{bv}(\boldsymbol{X})$ vanishes iff, for each pair $i \neq j$, X_i, conditional on the past of the entire remaining system *excluding* X_j (i.e. $\boldsymbol{X}_{[ij]}$), is independent of X_j. Where we know the existence of structural links $j \to i$ in the system, it may be appropriate to average the TEs only over these links. The Granger causality analogue of (4.48), termed *causal density*, was introduced in [301].

Another candidate, which we term *information transfer density* or *global transfer entropy* [24], is given by

$$ \begin{aligned} \mathbf{T}_{gl}(\boldsymbol{X}) &\equiv \frac{1}{n} \sum_i \mathbf{T}_{\boldsymbol{X}_{[i]}\to X_i}(t) \\ &= \frac{1}{n} \sum_i \left[\mathbf{H}(X_{i,t} \mid X_{i,t-1}) - \mathbf{H}(X_{i,t} \mid \boldsymbol{X}_{t-1})\right], \end{aligned} \tag{4.49} $$

which averages over the collective TE (Eqn. 4.20) into each variable i in the system. $\mathbf{T}_{gl}(\boldsymbol{X})$ vanishes iff each X_i, conditional on its own past, is independent of the past of the rest of the system, i.e. the past of $\boldsymbol{X}_{[i]}$.

Measures like $\mathbf{T}_{bv}(\boldsymbol{X})$ and $\mathbf{T}_{gl}(\boldsymbol{X})$ have been proposed in the neurosciences (Sect. 7.3) as reflecting a balance between *integration* and *segregation* of complex networks of dynamic processes [326, 301, 29]. For a highly segregated system, where elements behave near-independently, the measures will take on small values since there will be little feedback between processes. However for highly integrated systems the measures will also be expected to take on small values, since the system as a whole will have little information to add to that already contained in the past of a sub-process. Thus these measures will be highest for systems exhibiting a balance between integration and segregation, which has, in particular, been mooted as a hallmark of *consciousness* in the neuroscience literature [326, 302].

A further application of the measures is in the detection of *phase transitions* in large, complex ensembles of interacting elements. It has been established for a wide variety of model and real-world systems featuring order-disorder phase transitions (including spin systems, particle swarm systems, random Boolean networks (Sect. 5.3), neural systems (Sect. 7.3), financial markets (Chap. 6), and ecosystems) that mutual information between system elements tends to peak precisely *at* the phase transition. However, there is recent evidence [24] (see Sect. 5.2) that, at least for some systems, (global) information *flow* peaks on the *disordered* side of a transition, raising the possibility of *predicting* an imminent disorder $\rightarrow$ order transition in a system with slowly changing control parameters. This is of particular importance since, for many real-world systems (e.g. neural and financial market systems), order is associated with pathological dynamics (e.g. epileptic seizures and market crashes) whereas a healthy system features disordered dynamics.

4.7 Continuous-Time Processes

So far we have considered only processes where the time variable t is *discrete*. Here we ask how information transfer might be defined for processes with a *continuous* time variable. We remark that surprisingly little research appears to have been done in this area (but see e.g. [294]), with the notable exception of *point processes*, where Granger causality-like parametric measures have been proposed (see Sect. 7.3.2).

We consider jointly stochastic processes $X(t), Y(t)$ with a continuous (one-dimensional, real) time parameter t. One might then be tempted to define $\mathbf{T}_{Y\rightarrow X}(t)$ as $\lim_{dt\rightarrow 0} \mathbf{I}(X(t) : Y(t-dt) \,|\, X(t-dt))$. However there are problems with this: firstly, work in progress by the authors indicates that, for a class of multivariate *Ornstein–Uhlenbeck* (OU) processes [330, 80], which may be thought of as continuous-time analogues of VAR processes, in fact
$\mathbf{I}(X(t) : Y(t-dt) \,|\, X(t-dt)) \rightarrow 0$ as $dt \rightarrow 0$, although

$$\lim_{dt\rightarrow 0} \frac{1}{dt} \mathbf{I}(X(t) : Y(t-dt) \,|\, X(t-dt)) \tag{4.50}$$

generally approaches a non-zero finite value. This is perhaps not so surprising, and suggests that transfer entropy is best viewed as an information transfer *rate*—that

is, it measures the amount of information transferred *per unit time*. But a more serious problem with (4.50) is that in the limit $dt \to 0$ historical dependencies become instantaneous, whereas the joint processes may well feature feedback at *finite* time lags. This will be the case, for instance, for the vector OU process with distributed lags [27]

$$dU(t) = \left[\int_{s=0}^{\infty} A(t-s) \cdot U(s)\, ds \right] dt + dW(t), \tag{4.51}$$

where $W(t)$ is a *Wiener process* [80] (roughly, an integrated white noise process or *random walk* in continuous time) and the autoregression kernel $A(u)$ has finite mass in some interval away from zero.

We would, of course, like feedback at finite temporal lags to be taken into account. Now for a continuous-time stochastic process $U(t)$, the analogue of the history (4.10) of a discrete-time process U_t, $\boldsymbol{U}_{t-1}^{(k)} \equiv (U_{t-1}, \ldots, U_{t-k})$, is history-length τ past $\boldsymbol{U}^{(\tau)}(t) \equiv \{U(t-s) : 0 < s \leq \tau\}$. The problem here is that (even for finite τ) $\boldsymbol{U}^{(\tau)}(t)$ is an uncountably infinite set of random variables, and as such cannot be used naïvely in a putative expression like $\mathbf{I}\Big(X(t) : \boldsymbol{Y}^{(\upsilon)}(t) \,\Big|\, \boldsymbol{X}^{(\tau)}(t) \Big)$ for transfer entropy; moreover, such an expression would not in any case be operational for estimation from empirical data. A better approach is suggested by the practicality that, given a continuous-time process, empirically we will at best have computational access only to a finite sample of values; that is, a *discretisation in time* (or down-sampling) of the process. Thus for a continuous-time process $U(t)$, a small time increment dt and a finite history time lag τ we define [*cf.* (4.10)] the (finite) discretised history

$$\boldsymbol{U}^{(\tau)}(t; dt) \equiv U(t-dt), U(t-2dt), \ldots, U(t - [\tau/dt]\, dt), \tag{4.52}$$

where $[x]$ denotes rounding towards the nearest integer. We propose that the continuous-time transfer entropy with history (τ, υ) be defined as

Definition 4.10.

$$\mathbf{T}_{Y \to X}^{(\tau,\upsilon)}(t) \equiv \lim_{dt \to 0} \frac{1}{dt} \mathbf{I}\Big(X(t) : \boldsymbol{Y}^{(\upsilon)}(t; dt) \,\Big|\, \boldsymbol{X}^{(\tau)}(t; dt) \Big), \tag{4.53}$$

assuming the limit exists. Recent results [370, 27] indicate that this definition yields meaningful results, at least for processes of the form (4.51). However, further research is required to establish the class of processes for which Definition 4.10 is appropriate, or whether other types of continuous-time processes may require different treatment. We also remark that, empirically, care must be taken to choose a down-sampling time increment dt of size appropriate to the feedback time scales of the process. Recent results [27] indicate that (i) for optimal detection of information transfer at a given time lag, there is a "sweet spot" for dt slightly greater than the largest/typical causal lag time, and (ii) the ability to detect information transfer drops off exponentially for dt larger than the lag. As for discrete-time transfer entropy, parametric methods may often be preferable.

An important class of continuous-time stochastic processes are *point processes*, where discrete *events* occur at randomly distributed time intervals [71]. Point processes are of particular significance as models for neural *spike trains* in neuroscience; they require a rather specialised approach to definition and estimation of transfer entropy, and are discussed in detail in Sect. 7.3.2.

Chapter 5
Information Transfer in Canonical Systems

Having introduced the transfer entropy in Chap. 4, we now turn our attention for the remainder of the book to reviewing what this measure can tell us about various complex systems, and guiding the reader through these relevant applications of the measure.

In this chapter, we review applications of the transfer entropy to canonical systems—simple models of complex systems and networks which are widely used to study the nature of self-organisation and emergent behaviour. In particular, we describe the novel insights that TE produces when applied to:

- *Cellular automata* (Sect. 5.1)—that the well-known glider structures are the primary information transfer carriers;
- *Spin systems* (Sect. 5.2)—where maximisation of a multivariate TE near to a second-order phase transition may enable prediction of such upcoming transitions;
- *Complex networks*—including *random Boolean networks* (Sect. 5.3), where we observe a balance of computational capabilities near to criticality; and *small-world networks* (Sect. 5.4) which demonstrate that long links promote information transfer;
- *Flocking models* (Sect. 5.5)—revealing waves of co-ordinated motion as information cascades;
- *Synchronisation processes* (Sect. 5.6)—where TE activity is seen to be a key driver of synchronisation.

We will review the application to each of these model systems by describing:

1. What the model is and how it functions;
2. Why the concept of information transfer is important for that system;
3. How transfer entropy was measured;
4. What transfer entropy revealed.

This structure serves to didactically guide the reader's thinking about how to apply the transfer entropy to their own systems.

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_5

Crucially, the application of TE in each of these situations is more subtle than simply trying to use it as *the* elusive measure of complexity: we will see the use of the dynamics of information transfer in space and time, as well as its interaction with related quantities such as information storage, to build a description of a system. A crucial message from these applications is that:

Key Idea 28: *Using transfer entropy, even in these simple systems, requires some subtlety and thought about which information channels to measure and how to approach such measurement.*

These applications provide a foundation for the later chapters which examine the utilisation of TE in finance and economics (Chap. 6) and other application domains (Chap. 7) including neuroscience (Sect. 7.3).

5.1 Cellular Automata

We begin by reviewing the application of local transfer entropy to cellular automata (as presented in [193, 195, 191, 196, 198, 182]), revealing emergent "glider" structures as dominant information transfer entities.

What are cellular automata? As discussed earlier in Sect. 1.2.1, the canonical or original form of *cellular automata* (CAs) are discrete dynamical systems with an array of cells that synchronously update their value as a function of a fixed number of spatial neighbours using a uniform rule [362]. (There are variants to this form, e.g. with asynchronous connectivity.) The update rule is specified by listing the next value for a given cell as a function of each possible configuration of its neighbourhood in a rule table—see Table 5.1—and summarising this specification in a single number (known as a Wolfram number; see [362]). We focus here on elementary CAs (ECAs), which are 1D arrays of binary-valued cells with one neighbour on either side.

Although the behaviour of each individual cell in a CA is very simple, the (non-linear) interactions between all cells can lead to very intricate global behaviour, meaning CAs have become a classic example of self-organised complex dynamics. Of particular importance, CAs have been used to model real-world spatial dynamical processes, including fluid flow, earthquakes and biological pattern formation [225].

A particular reason for the interest in CAs within complex systems science is the emergence of self-organised coherent structures in the dynamics of certain rules. As described in Sect. 1.2.1, Wolfram [362] sought to classify the asymptotic behaviour of CA rules into four classes: I. homogeneous state; II. simple stable or periodic structures; III. chaotic aperiodic behaviour; and IV. complicated localised structures, some propagating. Class IV CAs (e.g. rules 110 and 54—see Fig. 5.1) are those with the aforementioned self-organised coherent structures; these complex rules are conjectured to lie at the "edge of chaos" (see e.g. Langton [174] and also

[251, 70, 228]) between the ordered rules of class I and II and the chaotic rules of class III (e.g. rule 22). Trying to automatically classify such rules has attracted much attention (e.g. [174, 366]) and indeed important questions over whether this is at all possible [119] – certainly (at this stage), there is no measure which can differentiate between ordered–complex–chaotic CA rules. Regardless, the idea of the classes does provide an interesting analogy (for discrete-state and time) to our knowledge of dynamical systems, and for our purposes, we are interested in the self-organised coherent structures as representative of dynamics of complex systems.

These emergent structures are known as *particles*, *gliders*, *blinkers* and *domains* (see e.g. Fig. 5.1). A domain is a set of background configurations in a CA, any of which will update to another configuration in the set in the absence of any disturbance. Particles are dynamic elements of coherent spatiotemporal structure, as disturbances or in contrast to the background domain. Gliders are regular particles, and blinkers are stationary gliders.[1] Several techniques exist to *filter* particles from background domains (e.g. [116, 117, 129, 130, 366, 136, 137, 303, 195, 196, 198]).

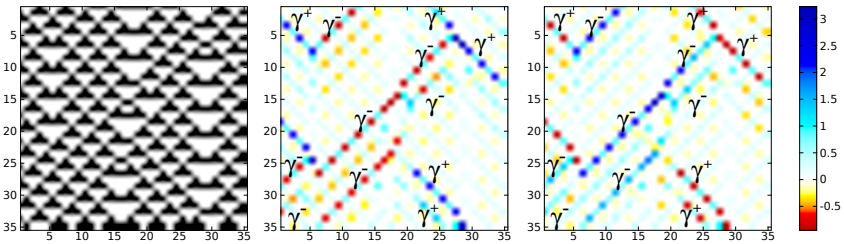
Why is information transfer important in CAs? These emergent structures have been quite important to studies of distributed computation in CAs, e.g. regarding universal computation (see [225]), and dynamics of intrinsic or other specific computation ([174, 129, 229]). Such studies typically discuss the computation in terms of three primitive functions of computation and their apparent analogues in CA dynamics [225, 174]:

- Blinkers as the basis of *information storage*, since they periodically repeat at a fixed location
- Particles as the basis of *information transfer*, since they communicate information about the dynamics of one spatial part of the CA to another part
- Collisions between these structures as *information modification*, since collision events combine and modify the local dynamical structures

Table 5.1 Rule table for ECA rule 110. The Wolfram rule number for this rule table is composed by taking the next cell value for each configuration, concatenating them into a binary code starting from the bottom of the rule table as the most significant bit (e.g. b01101110 = 110 here), and then forming the decimal rule number from that binary encoding.

Neighbourhood configuration for cell i at time n			Next cell value $x_{i,n+1}$ at time $n+1$
cell $x_{i-1,n}$ value (left)	cell $x_{i,n}$ value	cell $x_{i+1,n}$ value (right)	
0	0	0	**0**
0	0	1	**1**
0	1	0	**1**
0	1	1	**1**
1	0	0	**0**
1	0	1	**1**
1	1	0	**1**
1	1	1	**0**

[1] See formal definitions of these terms in [129].



(a) Raw CA values (b) $\mathbf{t}_{Y\to X}^{(k=16,\ell=1)}(t)$ right ($j=1$ channel) (c) $\mathbf{t}_{Y\to X}^{(k=16,\ell=1)}(t)$ left ($j=-1$ channel)

Fig. 5.1 Local transfer entropy in ECA rule 54 for the raw values in (a) (black for "1", white for "0"). 35 time steps are displayed for 35 cells, and time increases down the page for all CA plots. Local pairwise transfer entropy highlights gliders moving in the corresponding direction: (b) TE one cell to the right, and (c) TE one cell to the left per time step. Units are in bits—see scale at right-hand side. Reprinted with kind permission from Springer Science+Business Media (©holder), Figure No. 2 from [184]: J. T. Lizier, "Measuring the dynamics of information processing on a local scale in time and space", in M. Wibral, R. Vicente, and J. T. Lizier, editors, "Directed Information Measures in Neuroscience", Understanding Complex Systems, pages 161–193. Springer, Berlin/Heidelberg, 2014

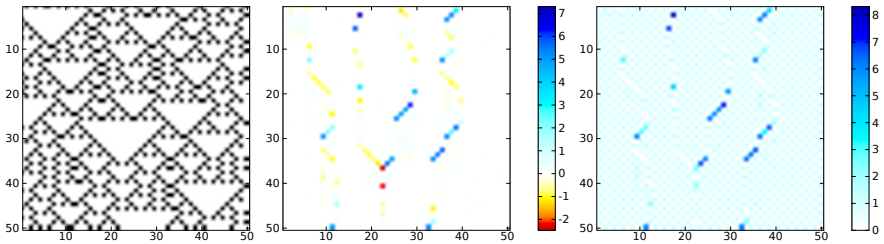
These analogies remained conjecture only, based on qualitative observation of CA dynamics. There was a strong need for a quantitative metric, which TE turned out to be.

How was transfer entropy measured in CAs? Observations for the relevant PDFs were taken over a short transient period for a large number of cells, and local transfer entropy $\mathbf{t}_{Y\to X}^{(k=16,\ell=1)}(t)$ (see Sect. 4.2.5) was computed for each time step t for each target cell X and for the two causal sources Y on either side of X (referred to as channels $j=1$ and -1 for transfer across one cell to the right or left). TE was computed by plugging in the discrete PDFs estimated from the data here.

The use of observations from only a short transient period aims to avoid non-stationarities in the data, and in particular to avoid sampling when the CA state has reached an attractor. This is because, once an attractor has been reached, each cell in the CA executes a periodic pattern which can only contain information storage (Sect. 4.2.2), leaving no scope for information transfer. From another perspective, we can say that the computation (of the attractor) by the CA would be completed at this point, leaving nothing for the TE to measure.

Sample results of this application are displayed for rules 54 and 18 in Fig. 5.1 and Fig. 5.2. The figures displayed here were produced using the open-source *Java Information Dynamics Toolkit* (JIDT) [183], which can be used in Matlab or Octave and Python as well as Java. All results can be reproduced using the Matlab or Octave script `TeBook2013.m` in the `demos/octave/CellularAutomata` example distributed with this toolkit.

What did transfer entropy reveal about CAs? The most important result from this application is that local transfer entropy is typically strongly positive at mov-



(a) Raw CA values

(b) $\mathbf{t}_{Y\to X}^{(k=16,\ell=1)}(t)$ left ($j=-1$ channel)

(c) $\mathbf{t}_{Y\to X|Z}^{(k=16,\ell=1)}(t)$ left ($j=-1$ channel)

Fig. 5.2 Local transfer entropy in ECA rule 18 for the raw values in (a) (black for "1", white for "0"). 50 time steps are displayed for 50 cells. (b) Local *pairwise* transfer entropy one cell to the left per time step; (c) Local *conditional* transfer entropy one cell to the left per time step. Units are in bits—see scales at right-hand side of figures. This figure first appeared in [184] Reprinted with kind permission from Springer Science+Business Media (©holder), Figure No. 3 from [184]: J. T. Lizier, "Measuring the dynamics of information processing on a local scale in time and space", in M. Wibral, R. Vicente, and J. T. Lizier, editors, "Directed Information Measures in Neuroscience", Understanding Complex Systems, pages 161–193. Springer, Berlin/Heidelberg, 2014

ing particles in comparison with blinkers and background domains [195]. This is when the local information transfer is measured at a particle in the same direction or channel j as the macroscopic motion of that particle. For example, see the highlighting of left (γ^-) and right (γ^+) moving gliders for rule 54 in Fig. 5.1b and Fig. 5.1c by transfer entropy to the left and right, respectively. Similarly, see the left moving sections of domain walls for rule 18 in Fig. 5.2b and Fig. 5.2c highlighted by transfer entropy to the left (TE to right omitted). In these examples, the source cell y_t which is in the particle at the previous time step t (be that the left or right neighbour, as relevant for that particular particle) is highly predictive about the next value of the target x_{t+1} (in the context of its past state $\mathbf{x}_t^{(k)}$, which is part of the background domain and cannot predict the particle being encountered). As such, we have $p(x_{t+1} \mid \mathbf{x}_t^{(k)}, y_t) > p(x_{t+1} \mid \mathbf{x}_t^{(k)})$, giving large positive values of $\mathbf{t}_{Y\to X}^{(k,\ell=1)}(t+1)$ via Eqn. 4.26. In contrast, in the domain the past state $\mathbf{x}_t^{(k)}$ is generally highly predictive of the next state—this executes strong information storage operations (see [198] and Sect. 4.2.2), but leaves little possibility for the source to add additional information transfer.

These results for local transfer entropy are particularly important because:

Key Result 1: *Local transfer entropy provides the first* quantitative *evidence that particles are the dominant information transfer agents in cellular automata. This result holds for related moving coherent spatiotemporal structures in other systems—see Sect. 5.5.*

An adequate embedded history length k is essential to properly capture the past *state* of the cell, and the results could not be observed with a value say of $k = 1$ (as discussed in further detail in [195, 337] and Sect. 4.2). Adequately capturing the past state of the cell can also be viewed as properly accounting for the role of information storage, or memory in the dynamics. Blinkers and regular background domains are dominant information storage entities, as identified via local active information storage [198]. We see the complementary nature of information storage and transfer here (as outlined in Sect. 4.2.2).

It is important to note that particles are not the only points with positive local transfer entropy (see discussion in [195]), though they are dominant.

Local information transfer is often found to be *negative* at moving particles, when measured in the orthogonal direction to macroscopic particle motion in space–time [195]. For example, see the measurement for TE to the left for the right-moving gliders in Fig. 5.1c. This is because the source Y here, being on the opposite side of the target to the incoming particle and therefore still part of the domain observed in the target's past, would suggest that this domain pattern would continue, which is *mis-informative*. (Recall that local MI and conditional MI values can be negative, where observing a source variable *reduces* an observer's expectation of the given actual outcome of the target variable—see Sect. 4.2.5.) That is to say, we have here $p(x_{t+1} \mid \mathbf{x}_t^{(k)}, y_t) < p(x_{t+1} \mid \mathbf{x}_t^{(k)})$, giving negative values of $\mathbf{t}_{Y \to X}^{(k=16,\ell=1)}(t+1)$ via Eqn. 4.26. These negative or mis-informative values are quite useful, since they imply that there is an extra feature in the dynamics that is unaccounted for in the past of the source and target alone.

We get complementary results if we condition out all the other interactions in determining the transfer entropy between two entities [195]. We know that conditioning can remove redundant information (see e.g. the ϕ_{par} CA rule analysed in [182]), but it can also include synergies (Sect. 3.2.3.1). For example, the background domain of rule 18 (see Fig. 5.2a) executes an exclusive OR operation (XOR) between left and right neighbours to determine the next state for a given cell. As we saw in Sect. 3.2.3.1, XOR is a highly synergistic operation. As such, we see that the pairwise (or "apparent") TE from one source only in Fig. 5.2b reports no information transfer in the background domain here; on the other hand the conditional[2] TE, which examines both sources, reveals strong higher-order transfer in Fig. 5.2c in capturing the synergies underpinning the dynamics here. We pick up on the complementary nature of these different TEs again in Sect. 5.3.

The differences between the concepts of information transfer (as captured by the transfer entropy) and causal effect were explored using local dynamics in CAs in [191]. The results may be *qualitatively* summarised as follows:[3]. A neighbour cell has the same direct causal effect on a target every time the same neighbourhood configuration (represented by a row in the rule table, see Table 5.1) occurs. Now since the same neighbourhood configurations which occur in the gliders also occur

[2] In fact, given that there is only one other causal source here, this conditional TE is a complete TE (see Sect. 4.2.3).

[3] Formally, this paper measured an intervention-based measure of causal information flow (from [11]) at every point on the CA.

in the background domains, then there is no difference in the level of direct causality that occurs in the gliders or the background domains. This is in contrast to transfer entropy, which as above was only comparatively high in gliders. This is because in dealing with *state* updates of the target, and in particular in separating information storage from transfer, the transfer entropy has a very different perspective to causal effect. A causal effect can be seen to serve either information storage (background domain) or transfer (glider), depending on the *context* of the past *state* of the target (as per Sect. 4.2.2). Again:

Key Result 2: *Neither a perspective of information transfer in computation nor causality in mechanics is more correct than the other—they both provide useful insights and are complementary.*

We emphasise also that only local measures reveal the richly structured spatiotemporal profiles here; this is not possible with the average measures (which only return a single number). Furthermore, the average values do not give so much as a hint towards the complexities of these local dynamics. The chaotic ECA rule 22 has no emergent self-organised particle structures (i.e. no coherent propagating structures), and yet has much larger average transfer entropy values than the complex rule 54 (0.19 versus 0.08 bits for each, respectively, in both left and right directions) [197]. That is:

Key Result 3: *High average TE does not imply the presence of coherent particle structures; only the local TE can reveal this.*

Similarly, while we do not yet have a single measure to differentiate between ordered–complex–chaotic CA rules (see Sect. 1.2.1), the local dynamics of emergent structures do seem more informative about the complexity of a given rule rather than any single measure (further explored in [197]).

Open Research Question 7: *Can local (or another variant of) transfer entropy be used to formally separate complex from ordered or chaotic dynamics?*

The above insights are in some sense similar to the results in the next two sections that TE does not necessarily peak directly at the critical point during a phase transition.

Finally, the reader may wonder whether there is some conservation of the information measured above in these systems. This is not the case, because the cellular automaton is not thermodynamically closed. Were we to examine both the CA and the underlying implementation of it (e.g. bit registers in a CPU), then this would

be the case. Investigations of relationships between transfer entropy and thermodynamic variables, as well as conservation properties, are currently underway; e.g. [275, 273], as discussed in Sect. 8.3.

5.2 Spin Models

We now want to consider another system, which looks very much like a cellular automaton at first glance. Spin models date back nearly a hundred years, and there is a huge literature surrounding them. But the new, exciting results obtained for transfer entropy are for one of the very simplest, the two-dimensional Ising model. There are many different spin models, but at the time of writing, transfer entropy has been computed for just one, the 2D Ising model, thus this will be the primary focus of our attention.

What are spin models? Just as a basic cellular automaton is a set of cells on a lattice, with a number of discrete states, mostly just binary, 0 and 1, as found in Sect. 1.2.1 and Sect. 5.1, so too is the Potts model. This is the eponymous model created by Renfrey Potts in the 1950s. If we make the states binary, effectively considering the spins as pointing up or down, then we have the Ising model [141], created by Ernst Ising.

The difference from 2D cellular automata arises in the way the cells are updated. In the CA this is based on a rule taking into account the states of the neighbours. The Ising model is more akin to a physical system. Spins interact with one another: their energy is lower when they are pointing in the same direction, as compared with when they are pointing in opposite directions. Thus at very low temperature, all the spins point in the same direction, but at very high temperatures, the directions of all the spins in the lattice are completely random.

Ising introduced his model as a theory of magnetic materials: an external magnetic field causes the spins to line up with the field, but in ferromagnetic, naturally magnetic, materials, when a sufficient number of spins line up the material becomes magnetic. But what might be a bit surprising is that the magnetisation does not gradually appear as the material is cooled down. There is a specific temperature, the Curie temperature, at which the magnetisation appears.

The Ising model shows the same sharp onset of magnetisation, and was solved exactly by Lars Onsager, for which he received the Nobel Prize in chemistry. The Curie temperature in fact marks a second-order phase transition, as discussed in Sect. 3.3.

Why is information transfer important in spin models? The Ising model subsequently became a canonical model for second-order phase transitions, and numerous studies have looked at the information-theoretic quantities. The mutual information was found to peak at the phase transition by Matsuda [218], and other more refined estimates have been made as recently as 2013 [176]. Until 2013, however, what happened to the transfer entropy was unknown. For the 2D Ising model, *no known physical quantity shows a peak away from the phase transition.* But finding

such a quantity could be enormously important: it would make impending second-order phase transitions predictable.

How was transfer entropy measured in spin models? There are different ways of updating spin systems, not something we have space to go into in great detail here. Barnett et al. [24] used Glauber updating [108]. Basically a spin is chosen at random as a candidate to be flipped at each time step. The flip goes ahead according to a probability, P_{flip}, dependent upon the energy, ΔE, required for the spin flip to occur:

$$P_{flip} = \frac{1}{1 + e^{\beta \Delta E}}, \tag{5.1}$$

where β is called the inverse temperature. At very large temperatures, as β tends to zero, P_{flip} tends to 0.5. If ΔE is positive, in other words if it requires energy input to make the spin flip, then as the temperature falls to zero, the denominator tends to infinity and the probability of the spinflip is zero.

Calculating mutual information and transfer entropy (as a function of temperature) follows the methods in Chap. 4, with the spin statistics computed from long time series after an initial settling period (in Barnett et al. [24] this was 10^5 updates after a settling period of 10^4 updates).

There is one last thing to consider. We can calculate the MI and TE between pairs of spins and average across all pairs—this gives the average *pairwise* MI and TE. Alternatively we can calculate the MI or TE between a spin and all the other spins (i.e. *collective* TE, see Eqn. 4.20), and average over all spins. This gives the *global* MI and TE (see Eqn. 4.49).

What did transfer entropy reveal about spin models? Neither the pairwise nor global mutual information show anything unusual with respect to temperature: they peak exactly at the phase transition (Curie point). Fig. 5.3 illustrates this and also shows that the pairwise transfer entropy peaks at the Curie point too. This maximisation of pairwise transfer entropy at the critical point is echoed by an empirical investigation of transfer entropy in the Ising model implemented on a human brain network [214].

But the global TE (i.e. the average collective TE) is completely different. It peaks on the disordered side of the transition. This result is very new (2013), and a full intuitive understanding has not yet been achieved.[4] But the implications are very far reaching. Second-order phase transitions abound in the natural and socio-economic world. Thus this result might enable us to predict upcoming transitions.

Before leaving this exciting possibility, we might mention a couple of caveats. Firstly, in practice we may not have enough data to get very accurate estimates of the global TE, which is far more data hungry than the other quantities. Simulation is limited only by computational grunt, but real data, from finance, ecology, wherever, may just not be sufficiently plentiful for accurate prediction.

[4] We note the corresponding results in [77] of TE being maximised in the disordered phase, when TE is estimated with an extension of the Rényi rather than Shannon entropy.

The second caveat is more subtle. The information flow before the transition might be predictive, but it might intrinsically mean that the system under observation is in some way committed, and cannot be diverted from the forthcoming tipping point. Think of heading for a collision in a large truck: it may be obvious that the collision is going to occur, but there is neither enough braking distance nor room to manoeuvre to actually avoid it. This is speculation, of course, and much further work needs to be done.

5.3 Random Boolean Networks

Random Boolean networks were introduced in Sect. 1.2.5 and here we examine the transfer entropy characteristics of the dynamics [194], showing that transfer peaks near an order–chaos phase transition as the network structure is altered.

What are Random Boolean networks? *Random Boolean networks* (RBNs) are a class of generic discrete dynamical network models. They are particularly important in Artificial Life, since they were proposed as models of gene regulatory networks by Kauffman [154] (and see [102]).

An RBN consists of N_G nodes in a directed *network* structure (the following concepts are shown in Fig. 5.4). The nodes take *Boolean* activity values, and update these in time as a function of the activity values of the nodes from which they have incoming links. The network structure is determined at *random*, subject to whether the in-degree[5] for each node is constant or stochastically determined given an average in-degree $\bar{K}$. Given the structure, the deterministic Boolean function or lookup table by which each node computes its next state from its parent nodes is also decided at *random* for each node individually, subject to a probability p of producing "1" outputs (p close to 1 or 0 gives low activity, close to 0.5 gives high activity). The nodes here are heterogeneous agents: there is no spatial pattern to the network structure (indeed there is no inherent concept of locality), nor do the nodes have the same update functions. In classical RBNs (CRBNs) considered here, the nodes all update their states synchronously as for CAs.[6]

RBNs are known to exhibit three distinct phases of dynamics, depending on their parameters: ordered, chaotic and critical. At relatively low connectivity (i.e. low degree $\bar{K}$) or activity (i.e. p close to 0 or 1), the network is in an ordered phase, characterised by high stability of states and strong convergence of similar macro states in state space. Alternatively, at relatively high connectivity and activity, the network is in a chaotic phase, characterised by low stability of states and divergence of similar macro states. In the critical phase (the *edge of chaos* [174]), there is per-

[5] The *degree* of a node in a network refers to the number of edges it has. Specifically, for a directed network structure, *in-degree* refers to the number of incoming edges to a node, whilst *out-degree* refers to the number of outgoing edges from a node.

[6] Asynchronous updating schemes are more biologically realistic [135], though all exhibit similar order–chaos phase transitions [104, 103].

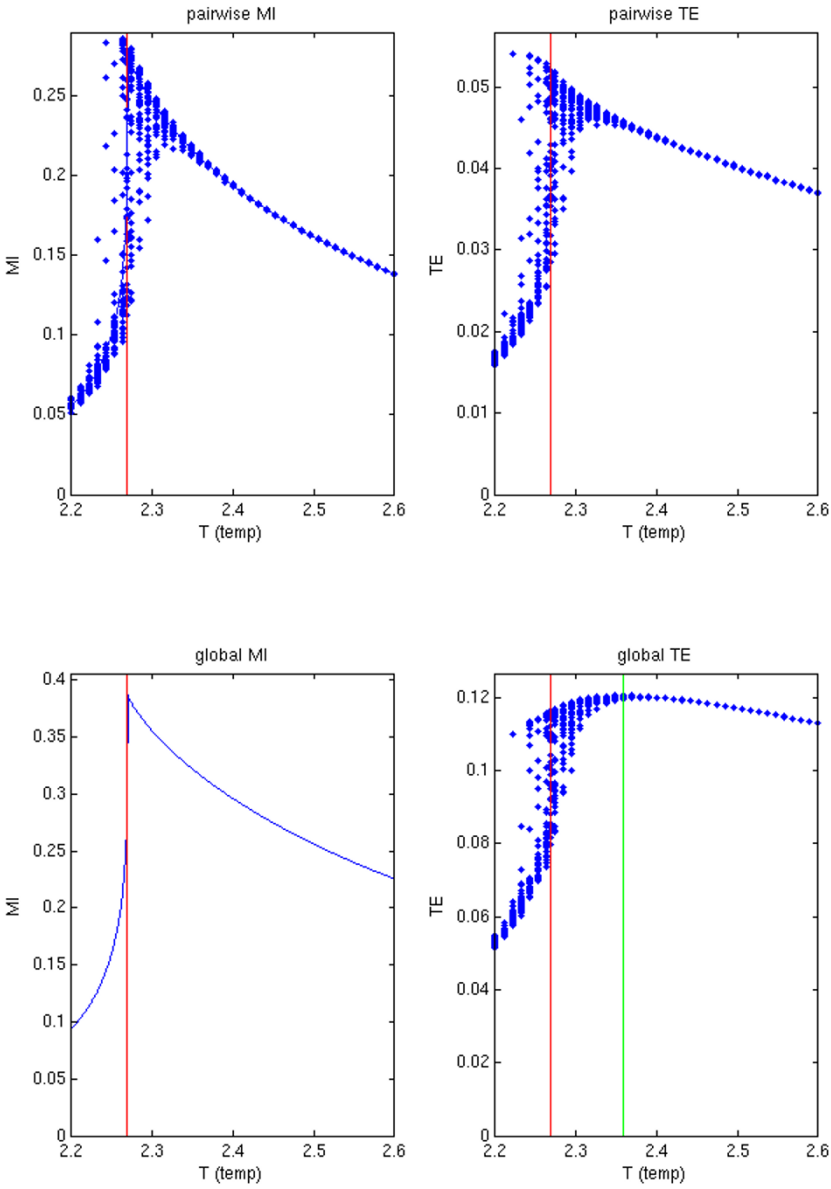


Fig. 5.3 Mutual information and transfer entropy for the Ising model. The red vertical line denotes the phase transition (Curie temperature). The green line shows the position of the peak for the global transfer entropy (after [24])

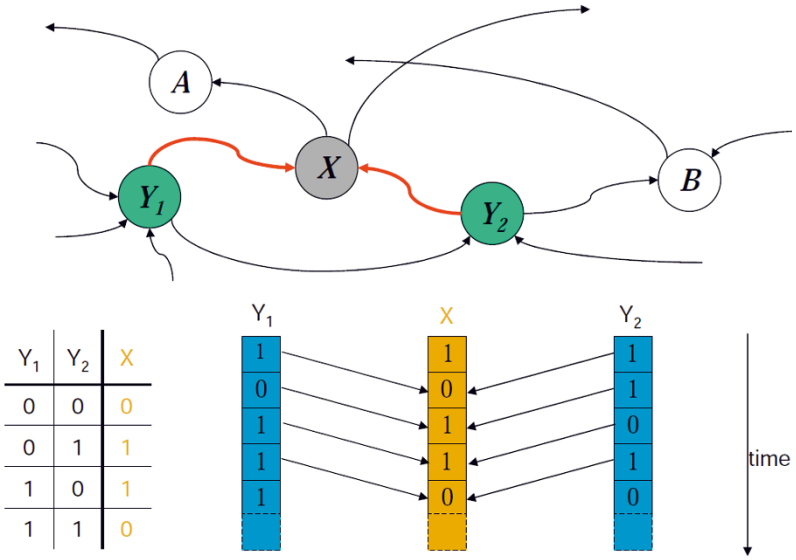


Fig. 5.4 A portion of an example RBN, showing the randomly determined update rule for node X as a function of its input nodes Y_1 and Y_2, and the synchronous time-series updates of these nodes (after [190])

colation in nodes remaining static or updating their values, and uncertainty in the convergence or divergence of similar macro states.

Why is information transfer in RBNs important? Much has been speculated on the possibility that gene regulatory and other biological networks function in (or evolve to) the critical regime (see [102]). It has been suggested that computation occurs more naturally with the balance of order and chaos there, i.e. at the *edge of chaos* [174], with maximisation of computational properties there [154]. There are conflicting interpretations however on such computation. Langton [174] suggests that an intermediate level of information propagation and storage gives rise to complex computation in critical dynamics, with too much of either decaying the computational capability. Others argue for maximisation of information transfer in this regime, e.g. [280, 309].

Recently, Ribeiro et al. [283] measured mutual information in the states of random node pairs as a function of connectivity in the network, and Rämö et al. [280] measured the uncertainty (entropy) in the size of perturbation avalanches as a function of an order parameter. Both studies found maximisations near the critical point, and claimed that their results imply maximisation of information propagation in this regime. The results are certainly interesting, but do not directly measure the concept of dynamic, directed information transfer.

How was transfer entropy measured in RBNs? Lizier et al. [194] sought to improve on these attempts to measure information transfer, by using TE. The main goal was to characterise the average information dynamics in an RBN as a function

of the average in-degree $\bar{K}$, through the order–chaos phase transition with respect to that parameter. To facilitate empirical measurements, many RBNs were generated with $N_G = 250$ nodes, balanced activity $p = 0.5$, for several values of $\bar{K}$, with their dynamics generated by the RBNLab software [101]. Observations for the relevant PDFs for $\mathbf{T}_{Y \to X}$ for each directed edge $Y \to X$ in each given network were accumulated over short transient periods[7] from many random initial conditions for the network. Each $\mathbf{T}_{Y \to X}$ was computed by plugging in the discrete PDFs estimated from the data here. The average TE was taken across all directed edges in all networks for the given $\bar{K}$ (i.e. as per Eqn. 4.47, but only where edges existed). In a similar fashion, Lizier et al. also measured the complete (or pairwise-conditional) transfer entropy (Eqn. 4.19), and the complementary measure of active information storage (Eqn. 4.14), averaged across all edges and nodes, respectively.

What did transfer entropy reveal about RBNs? The measurements of information dynamics in [194]—see Fig. 5.5—demonstrated that:

Key Result 4: *The ordered phase in RBNs is dominated by information storage (information already in nodes dominates their next states; the chaotic phase is dominated by information transfer (information from incoming links, in the context of the nodes' past, dominates their next states); there appears to be a balance between these operations near the critical phase.*

This correlates very well with Langton's conjectures regarding computational properties in complex systems [174]. The results also correlated well with a similar study of information storage and transfer through an order–chaos phase transition in recurrent neural networks [41].

More specifically, Lizier et al. investigated the decomposition of the information from incoming links. The pairwise TE (also referred to as apparent, single-source or bivariate TE) rises to a maximum value close to the critical phase, then falls away as the dynamics became more chaotic. The complete TE (Eqn. 4.19), on the other hand continues to increase into the chaotic phase. This implies that as the activity level increases, single sources are first observed to be having large predictive effects on the targets, allowing propagation of coherent effects near the critical regime. However these coherent single-source effects are swamped by the increasing level of interaction in the network, as activity increases with $\bar{K}$ in the chaotic regime. Prediction of the state transitions of targets becomes more efficient in examining multiple source nodes; in other words, information shifts more into higher-order transfer entropy terms such as the complete TE.

Interestingly, these results help to resolve the conjecture around information transfer through the phase transition, by using these two complementary information transfer measurements. Together, they clarify that it is single-source coherent transfer which peaks near the critical regime, while higher-order multivariate trans-

[7] Where the network is still *computing* its attractor; see Sect. 5.1 and [194].

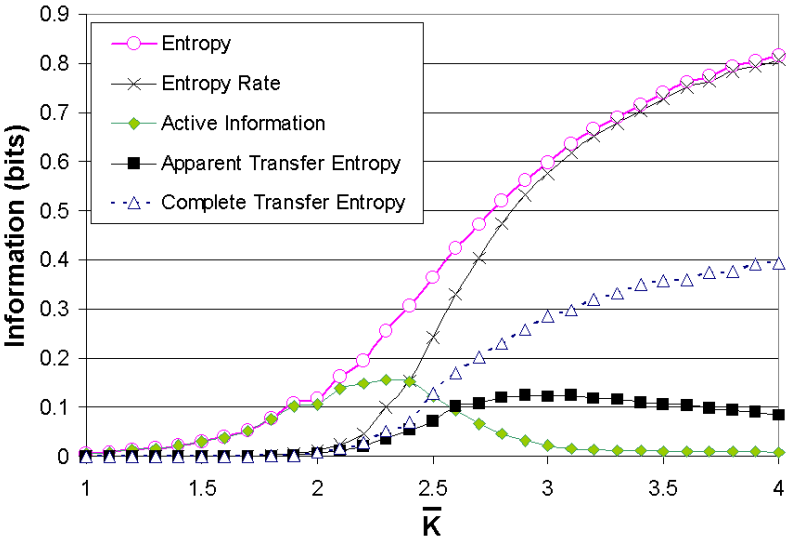


Fig. 5.5 Average information dynamics versus average connectivity $\bar{K}$ for RBNs of size $N_G = 250$. Plotted here are the average single node entropy $\mathbf{H}(X)$, entropy rate $\mathbf{H}'_X(t)$ (see Eqn. 4.15), active information storage (Eqn. 4.14), and average pairwise transfer entropy and complete transfer entropy on each directed edge. Note that entropy rate represents the sum of all orders of transfer entropy terms here (see [194, 196], Sect. 4.2.2 and Sect. 4.2.3). Error bars (omitted) are on the scale of the data points. Reprinted from [194]: J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "The information dynamics of phase transitions in random Boolean networks", in S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, "Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems, (ALife XI)", Winchester, UK, pages 374–381, © 2008 Massachusetts Institute of Technology, published by the MIT Press, Cambridge, MA

fer continues to increase into the chaotic regime. Considering these results, along with those for CAs in Sect. 5.1, we see that:

Key Result 5: *Conditional and pairwise transfer entropies reveal different aspects of the dynamics of a system—neither is more correct than the other; they are both useful and complementary.*

Comparing with Sect. 5.2, of course, the higher-order multivariate transfer terms themselves peak and decay in the chaotic regime if we have stochastic dynamics which cannibalise the predictability of sources. This is what we see with the Ising model, but it does not occur here since the dynamics are deterministic. Further, there are some indications that the peak for the pairwise TE would move towards criticality (as per the Ising model results) as the system size $N_G \to \infty$ (e.g. see [283]).

5.4 Small-World Networks

What are small-world networks? The small-world network model was proposed by Watts and Strogatz [347] to explore the "six degrees of separation" phenomenon [222] in *complex networks*; that is, how apparently highly clustered networks such as regular lattices can have small average path lengths similar to random graphs. The model has become one of the most influential concepts in complex systems science (e.g. see [152, 175, 204, 322, 369]), because of its ability to explain this phenomenon in a simple fashion, as well as the prevalence of small-world-type networks in both naturally occurring and man-made networks (e.g. in power grid networks and neural networks).

The model specifies how to tune networks from ordered, lattice-like structures, through small-world networks, and finally to fully random topologies. To start, construct a regular lattice network, such as a ring lattice with N_G nodes where each node is connected to $\bar{K}$ nearest neighbours. Then, rewire each edge in the network with a probability γ (moving one end of the edge to another node selected at random). This induces a phase transition in the network (seemingly of first order; see [241, 6]) between being completely ordered at one extreme ($\gamma = 0$) and completely random at the other ($\gamma = 1$). To quantify the effect of these random rewirings, Watts and Strogatz [347] suggest measuring the average *clustering coefficient* $C(\gamma)$ across all nodes, and the average *path length* $L(\gamma)$ between all node pairs. The clustering coefficient C_i for a node i is defined as the proportion of pairs of neighbours j and k of i that have an edge. For undirected, unweighted networks, we write $i \sim j$ to indicate an edge between nodes i and j, and write the degree of i as d_i; the clustering coefficient is then defined as:

$$C = E\{C_i\}, \tag{5.2}$$

$$C_i = \frac{1}{d_i(d_i-1)/2} \sum_{j,j\sim i} \sum_{k,k\sim i} \begin{cases} 1 & j \sim k \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

The path length L_{ij} for a node pair i and j is the length of the shortest path of edges connecting i and j, so we have

$$L = E\{L_{ij}\}. \tag{5.4}$$

In ordered, lattice-like networks, the high proportion of local links in these spatially embedded networks means that $C(0)$ is high, while $L(0)$ is also large. In contrast, the lack of spatial structure in random networks means that both $C(1)$ and $L(1)$ are small. Intermediate values of γ provide very interesting results however. Watts and Strogatz [347] demonstrated that there is a significant range of γ values for which the networks exhibit high clustering $C(\gamma)$ (comparable to fully ordered networks) and small average path length $L(\gamma)$ (comparable to fully random networks). Networks in this intermediate range are labelled *small-world networks*. The convergence of these properties occurs because even a small level of randomisation of the edges

creates "short cuts" across the network, which drop $L(\gamma)$ very quickly with respect to γ. Since this occurs with a relatively small level of randomisation though, the clustering $C(\gamma)$ remains relatively high. This means that one can reach a given node very quickly from any other node, even though the network "feels" very clustered.

Why is information transfer in small-world networks important? The structure and generation of small-world networks are well understood, yet questions remain over what the dynamic computational properties are that make them so useful in nature. Indeed, this issue pertains to the wider field of network science: network structure has attracted much attention, while time-series dynamics remain "much less well understood" [226]. Understanding the dynamics on networks is of vital importance: certainly structure gives rise to time-series dynamics on networks, but dynamics represent the specific action of a network, and only they can answer why a network is actually useful.

While much work regarding time-series dynamics has focussed on state-space trajectories and damage spreading, Mitchell [226] suggests that "the main challenge is understanding the dynamics of the propagation of information ... in networks, and how these networks process such information." This comment very nicely summarises the speculation regarding *computational properties* of networks, underlining why quantitative studies of transfer entropy on complex networks will be important.

Indeed, much of this speculation has focussed on the computational properties of small-world networks. Watts and Strogatz [347] themselves claimed that small-world topologies impart both "enhanced signal-propagation speed" and "computational power". Similarly, Latora and Marchiori [175] suggest that small-world networks are prevalent in nature because they balance local efficiency and global efficiency of information transport, suggested to be supported by local structure and long links, respectively. Tassier and Menczer [322] infer that small-world networks emerge in a model of evolutionary labour markets as a means to transfer information, while Katare and West [152] claim that small-world structures have "maximum capability to store, process, and transfer information". Despite such interest, the information storage and transfer capabilities of these networks (in particular using transfer entropy) had not been directly measured.

How was transfer entropy measured in small-world networks? Lizier et al. [190] aimed to directly investigate whether small-world networks do indeed maximise information transfer capability, using TE. The network structures under investigation were imbued with time-series dynamics by assigning random Boolean functions to their nodes, which amounts to combining RBNs with small-world network structures. RBN dynamics were selected for their ability to generate a wide range of dynamics, which was very suitable for an ensemble study of the dynamic properties of small-world networks. The main goal was to characterise the average information dynamics in these networks as a function of (the small-world parameter) rewiring probability γ, average in-degree or connectivity $\bar{K}$, and activity level r in the Boolean dynamics.

The ensemble study was conducted in the same manner as described for general RBNs in Sect. 5.3, with the addition of the extra parameter γ, with networks of size

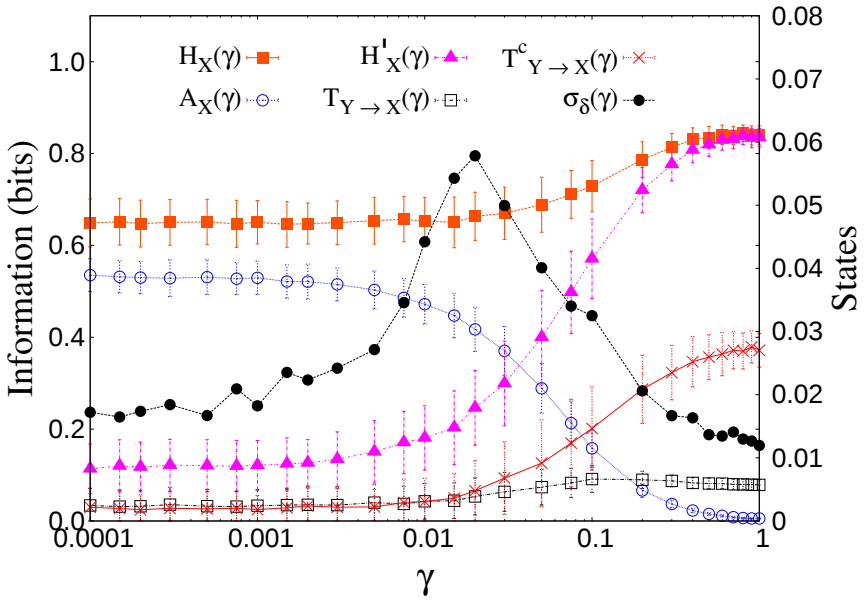
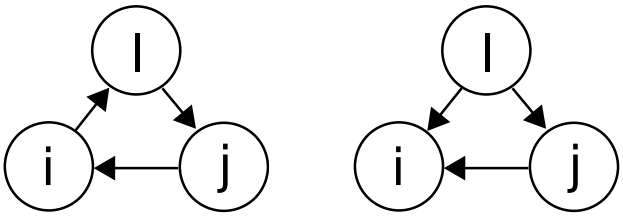


Fig. 5.6 Information measures versus γ, for networks with $\bar{K} = 4$ and $r = 0.36$ (after [190]). Information measures are in bits and plotted against the left y-axis: entropy, $\mathbf{H}(X)$; active information storage, $\mathbf{A}_X^{(k=14)}$; entropy rate, $\mathbf{H}'_X$; pairwise TE, $\mathbf{T}_{Y\to X}^{(k=14)}$; complete TE, ${}^c\mathbf{T}_{Y\to X}^{(k=14)}$. Note that the entropy rate here represents the sum of all orders of transfer entropy terms $H_{\mu X}$ (see Sect. 4.2.2). A measure of complexity in dynamics, σ_δ (a standard deviation of perturbation avalanche sizes; see [190] for full definition), is plotted against the right y-axis, with its peak indicating the critical regime of dynamics here—we have a subcritical regime to the left of this peak, and supercritical to the right. Error bars indicate the *standard deviation* of the values across the 250 sampled networks. (The standard error of the mean is too small to be visible)

$N_G = 264$. A key point is that the undirected regular networks which are rewired with probability γ were converted into directed networks, with each undirected link becoming two directed links, subjected separately to rewiring. Rewiring of both the source and target of links was investigated, giving mostly similar results, though only source rewiring is considered here (maintaining the in-degree of each node).

What did transfer entropy reveal about small-world networks? The measurements of information dynamics in regular–small-world–random networks in [190]—see Fig. 5.6—demonstrated order–chaos phase transitions of a similar nature to those for traditional RBNs, but with additional structure due to the rewiring parameter γ (see Fig. 5.6):

Key Result 6: *Networks with low levels of rewiring γ (more regular structure) and small activity r exhibit more ordered dynamics which is dominated by information storage, while networks with higher levels of rewiring γ (more*



(a) Three-node feedback loop (b) Three- node feedforward loop

Fig. 5.7 Motifs implicated in calculation of information storage at node *i* include *directed feedback cycles* and *feedforward loop motifs* (loops of length 3 shown for both types). This figure first appeared in [185] and is © American Physical Society, and is reprinted with permission

random structure) and higher activity r exhibit more chaotic dynamics which is dominated by information transfer.

The results suggest first that information storage is strongly supported by the clustered structure in regular or locally connected networks (with a significant correlation between storage and clustering coefficient reported). This would be expected with neighbours sharing common information here: in feedback loops for example (see Fig. 5.7), one can easily imagine information cycling around the loop to recur in the same node at multiple time steps. Similarly, the clustered structure here serves to segregate nodes and therefore limits the availability of novel or surprising information to be transferred to them. Further evidence for the dominance of active information storage dynamics in regular networks was provided by analytic findings that *feedback* and *feedforward loop motifs* (see Fig. 5.7—these are particularly prevalent in locally connected networks) directly support information storage operations for coupled Gaussian dynamics [185]. The results also suggest that information transfer is strongly supported by the introduction of long links as the network is randomised (with a significant anti-correlation between transfer and average path length). Again, one would expect long links to provide new information to target nodes that they would not receive from spatially close sources.

Further, the results indicate that the structural crossover between regular and random structure is mirrored by a crossover in dynamics from information storage to transfer being dominant. While the precise location of the critical regime with respect to γ varies with r and $\bar{K}$, in general small-world networks exhibit a propensity to balance information storage and transfer in their dynamics. These results (from transfer entropy and other measures) could be seen to add evidence for findings that:

Key Result 7: *Small-world networks hold computational advantages over regular or random network structures, in supporting both intrinsic information storage and transfer operations.*

In conjunction with structural constraints such as wiring costs, such computational advantages could be a driver in the emergence of small-world structures in networks in nature, in particular in brain networks.

Finally, in a similar fashion to the transition in RBNs, the results in Fig. 5.6 indicated that pairwise transfer entropy peaks on the chaotic side of the phase transition in dynamics (driven either by network randomisation γ or activity r). As the dynamics were driven further into the chaotic regime, the information composition again moved towards higher-order transfer entropy terms (see Sect. 4.2.3), as can be seen with the complete TE in Fig. 5.6.[8]

5.5 Swarming Models

What is swarming behaviour? Swarm behaviour refers to collective behaviour exhibited in movement by a group of animals [181, 257], or indeed artificial systems such as robots [43]. For specific types of animals, it is also known as flocking (birds), schooling (fish—see Fig. 5.8) or herding (buffalo). Such behaviour is thought to provide biological advantages in terms of protection from predators, mate choices, foraging etc. [52, 107]. As described in Sect. 1.2.6, intricate large-scale patterns and structures can emerge from swarm behaviour, including cascades of small perturbations travelling across a swarm in a wave-like manner [266] (e.g. waves of turning motion [278]), splitting and reforming of groups [258], group avoidance of obstacles and vortex-like "milling" behaviour where individuals rotate around an empty core [257].

Realistic simulation of swarm behaviour can be generated using three simple rules for the behaviour of each individual in the swarm, originally captured in Reynolds' *boids* model [282]:

- *Separation*—move to avoid collisions with other local individuals
- *Alignment*—move towards the average heading of other local individuals
- *Cohesion*—move towards the average position of other local individuals

The parameters of these models (e.g. sensory ranges or radii of local interactions), with some variations, can be tailored to simulate behaviour of many different species [153, 224].

Why is information transfer in swarming important? Concepts of information flow have often been used to qualitatively describe the dynamics of swarms.

[8] Again, unlike the Ising model, there is no stochasticity to reduce the higher-order TE terms as we move into the chaotic regime.

Fig. 5.8 Schooling groups of predator and prey fish. Schooling in fish produces apparent information cascades [67, 39], e.g. in handling predator avoidance by the school. This figure "Moofushi Kandu fish.jpg" is copyright by Bruno de Giusti, used under Creative Commons CC-BY-SA-2.5-IT [75]

Several authors argue for a relationship between critical swarming behaviour and some type of optimisation of information transfer. Couzin [66] invokes information transfer to interpret effective flocking behaviour occurring only at intermediate sensory ranges between individuals, suggesting that too short a sensory range does not allow enough information transfer to form cohesive groups, while too large a range permits rampant spreading of irrelevant information which erodes group cohesion. From another perspective, Vanni et al. [333] suggest that long-range correlations induced by criticality in the swarm enable efficient information transmission across the swarm.

More specifically, the aforementioned cascades of perturbations which cross swarms in a wave-like manner [266, 278] have been conjectured to embody information transfer, being labelled as *information cascades* [67, 39]. This seems quite reasonable as one can easily interpret information about turning to avoid obstacles or predators as being communicated in this fashion (see Fig. 5.8). Indeed, it has been observed that these mechanisms seem to allow information to be transferred over long ranges, and at faster speeds than incoming predators travel, perhaps conveying an evolutionary advantage [127, 67]. With that said, it has also been observed that such sensitivity to cascades comes at a price of fragility of co-ordinated behaviour [39] and susceptibility to noise or *false alarms* [107, 97, 67]. The importance of such coherent wave structures is underlined in that they are observed in many other animal groups, e.g. giant honeybees [151] and Emperor penguins [371], and indeed have analogies in other systems including perturbation waves in protein networks [5], the dynamic opening and closing of stomatal apertures in plants [260] and in gliders in cellular automata (as discussed in Sect. 5.1).

Information is a crucial currency for animals, with biological information processing important from both a behavioural and evolutionary perspective [263, 72].

This is the case at several levels, starting at the level of genetic networks as discussed later in Sect. 7.4, coming up to information processing in the brain as discussed in Sect. 7.3, and then information processing at the social level, for example, by swarms. Katz et al. [153] draw attention to exploring how animals in swarms integrate information from widely disparate sources and how this translates into higher-order computational capabilities. Couzin [66] concludes that swarms "may adapt to compute 'the right thing' in different contexts", where the right thing may be the optimal escape manoeuvre from a predator or the route around an obstacle, and Couzin et al. describe information cascades as being part of information processing in fish schools along with "collective memory" [67]. While swarms provide a simple model of dynamic interactions, information processing in social systems is not restricted only to swarms, being observed in human systems too of course, for example the existence of memory in collective editing on Wikipedia [76].

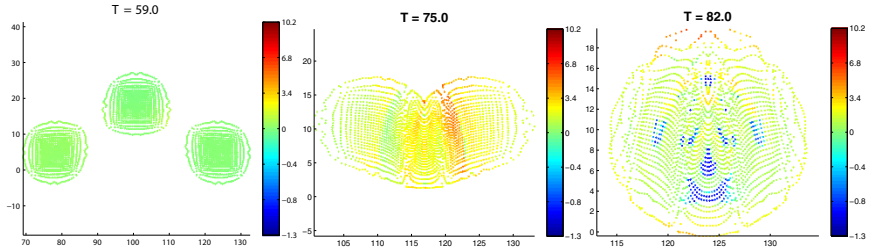
How was transfer entropy measured in a swarming model? Wang et al. measured transfer entropy in a three-zone swarming model using time series of positions and velocities of the agents in the swarm [344, 345]. Two important modifications were made to the usual calculation of transfer entropy.

First, Wang et al. noted that, unlike the homogeneous coupled pairs of variables in CAs or the heterogeneous but fixed coupled pairs in RBNs, "swarm computation is *amorphous*, with neither homogeneous computational structure across agents, nor with fixed computational relationships between heterogeneous causal pairs" [344]. That is, one should not simply compute TE on the whole time series for two interacting agents in a swarm, because they are highly unlikely to be within each other's relevant radii of interaction for many of those time steps, impairing the meaning of the measure. Instead, Wang et al. recommended taking advantage of the homogeneous functionality across agents in the swarming model, and using observations from *every* (within radii) pairwise causal interaction in computing the relevant PDFs for a TE calculation. This would then represent TE for a typical causally interacting pair in the swarm.

Second, Wang et al. used *relative* position and heading variables (with respect to those of the previous state of the information target agent) in the TE calculation instead of the *absolute* position and heading of the agents. The idea here is that the causal interactions depend only on these relative rather than absolute variables, and using these allows more observations for and better representation of the PDFs describing the interactions. The calculation performed was a conditional TE (Eqn. 4.18), conditioning on the absolute speed of the information target, since this has the potential to modulate the interaction.

TE was computed via kernel estimation, with kernel widths of 0.23 standard deviations for each variable. Local (conditional) TEs (see Sect. 4.2.5) were then computed for every causal interaction between locally connected agents in the swarm, with [344] analysing the temporal dynamics of these TE values averaged over all connected pairs at each given time step, and [345] examining the full spatiotemporal dynamics of local TE in the swarm.

What did transfer entropy reveal about swarming? The initial examination of the averages across the swarm in [344] indicated that TE peaked in the swarm during



(a) t=59 (b) t=75 (c) t=82

Fig. 5.9 Local transfer entropy at each agent in a swarm at several time steps as three separate swarms merge. The x–y coordinates of each agent in the swarm are indicated by the axes; the colour of each agent represents its local TE (averaged over TE contributions from each source to that agent)—red represents positive local TE, while blue is negative. These figures were first published in [345], and are copyright to the authors of that paper; the figures are re-used under the Creative Commons attribution licence. A video showing the local TE during this merge in more fine-grained detail is available on YouTube at http://youtu.be/vwfhijoq4cs, with further videos available in the playlist http://goo.gl/3QbQE8

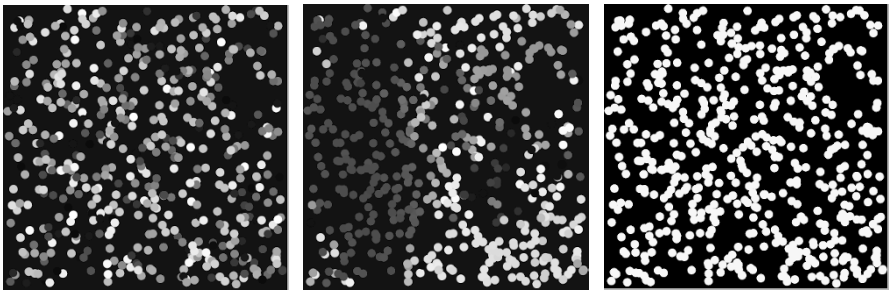
transition stages of collective behaviour, e.g. with two swarm fragments merging. The results verified the concept that "swarming dynamics can be interpreted as a type of distributed computation", i.e. with the agents transferring information via their relative positions and headings in order to compute their next stable group configuration.

Perhaps more importantly, the investigation of local transfer entropy in [345] revealed interesting patterns of space–time information dynamics in the swarm. These patterns are displayed for example in Fig. 5.9 (with links to more detailed videos in the figure caption). Crucially:

Key Result 8: *Wang et al. provided the first quantification of coherent information cascades in the swarm as waves of large, coherent information transfer.*

See for example the (red) wave of strongly positive local TE on the right side of Fig. 5.9b. This finding was particularly important as it provided the first direct information-theoretic evidence for such information cascades, which had been conjectured previously as described in Sect. 5.5.

The information cascades included wavefronts of both positive transfer entropy—indicating waves of strong influence, and negative transfer entropy (see blue-coloured waves in Fig. 5.9c)—indicating misinformation, perhaps as ineffectual influence, where the target agents respond in a different way to the source agents than would usually be predicted. The latter may occur where multiple sources are



(a) Unsynchronised (b) Emerging synchronisation (c) Synchronised

Fig. 5.10 Snapshots during a synchronisation process emerging from locally coupled oscillators. Grey-scale indicates the phase of each oscillator. We see the system move from: (a) an unsynchronised state, through (b) emerging synchronised pockets, to (c) a fully synchronised state. (Figures generated using the NetLogo model "Sync model" [53])

influencing a target agent in different ways, and the behaviour of the target is not predictable from a single source in isolation.

5.6 Synchronisation Processes

What are synchronisation phenomena? "Synchronisation phenomena" refers to the ability of a group of weakly interacting oscillators to mutually entrain [262], approaching a configuration where their individual states either coincide or have a constant phase difference. A simulated example is shown in Fig. 5.10. As described in Sect. 1.2.3, synchronisation phenomena have been independently observed in many different fields, including in swarms of flashing fireflies, clusters of pacemaker cells in the human heart and in electrons in superconductors [317, 262]. Researchers also study synchronisation in its own right, rather than in the subject-specific domains mentioned above, and have made synchrony a particularly important concept within complex systems science.

There are two aspects to consider regarding models of synchronisation phenomena: the *dynamics* of the individual units, and the connection *structure* between them. Arguably, the Kuramoto model [169, 170] has been the most influential model of dynamics of individual units.[9] The Kuramoto model considers a set of P connected oscillators X with fixed individual natural frequencies ω_X, and non-linear couplings between individuals either speeding up or slowing down their oscillations in a fashion to pull their phases closer together. The rate of change $\frac{d\theta_X(t)}{dt}$ of the

[9] Analysis of linear dynamics, including eigenvalue analysis, has also yielded many important insights into synchronisation (e.g. [8, 147]), and can be viewed as examining weakly coupled near-linear dynamics around a synchronised attractor state in a non-linear system (e.g. a system of Kuramoto oscillators).

phase $\theta_X(t)$ of oscillator X is defined as a function of the adjacency matrix A (where $A_{XY} = 1$ for a connection $Y \to X$ and 0 otherwise), *coupling strength* κ and the phase differences $\theta_Y(t) - \theta_X(t)$ from the oscillators Y to which it is coupled:

$$\frac{d\theta_X(t)}{dt} = \omega_X + \kappa \sum_{Y=1}^{P} A_{XY} \sin\left(\theta_Y(t) - \theta_X(t)\right). \tag{5.5}$$

For most connection structures the Kuramoto model exhibits a phase transition (which may be first or second order [30]) from unsynchronised to synchronised behaviour as the *coupling strength* κ between the elements is increased. Yet while the behaviour of some special structures (e.g. fully connected systems), or under linearisations or mean-field assumptions, can be identified by tractable analytic solutions [165, 31, 9], in general other topologies yield intricate behaviour, and it is non-trivial to determine whether a given non-linear system will support synchronisation without running a full simulation.

Empirically, the coherence of the system is generally measured using an *order parameter* (see Sect. 3.3), defined here as the magnitude of the average of all oscillators' phase vectors—the parameter is 0 for a uniform distribution of phases, and reaches 1 under complete synchronisation. A phase transition in this parameter may be observed for most connection structures as the coupling strength between units (as a control parameter) is increased.

Why is information transfer in synchronisation processes important? The idea that oscillators are communicating information about their states, and using such communication to settle on a shared information state, is quite intuitive. Indeed, Ceguerra et al. [54] interpret the synchronisation process as a distributed computation of whether or not synchronisation will occur and what the synchronised state will be, and discuss the information transfer and storage operations underpinning this process. This information processing perspective was later echoed by Bollt [42]. Ceguerra et al. [54] argue that the examination of transfer entropy in the synchronisation process has the potential to generate new insights in this fashion, in particular in revealing time-series dynamics of the information processing involved, and providing findings comparable to dynamics in other systems.

Yet what transfer entropy may reveal is not immediately clear. At first glance, stronger coupling may imply stronger transfer, but this certainly is not always the case in other systems (see e.g. [11] or Fig. 4.1). Furthermore, we also expect that transfer should be zero once synchronisation is achieved (since the oscillators' futures are then predictable from their past alone), and it is unclear how to resolve these potentially conflicting intuitions.

Similar impetus is provided by the *communication through coherence* hypothesis in neuroscience [94]. This hypothesis suggests that "only coherently oscillating neuronal groups can interact effectively because their communication windows for input and for output are open at the same times" [94]. In other words, the suggestion is that synchronisation influences the interaction between neural groups. In investigating this claim, Buehlmann and Deco [50] found that transfer entropy increased with synchronisation in the gamma frequency band. This is a complicated result—

we would not expect transfer to take place in the synchronised band alone—and studies of fundamentally simpler synchronising systems are required to understand the findings.

How was transfer entropy measured in a synchronisation process? Bollt [42] measured TE during a synchronisation process, but only for a pair of coupled oscillators, noting the impetus to go on to analyse TE during synchronisation in a complex network.

Ceguerra et al. [54] earlier made a more comprehensive study of transfer entropy in the dynamics of the synchronisation process in complex networks using the Kuramoto model. They constructed networks with structure derived from wireless sensor networks, with $N_G = 100$ nodes scattered randomly on a grid, and connectivity between all neighbours within a fixed radius of each other. The nodes began each simulation run from random phases, with subsequent Kuramoto updates. Networks were simulated with a range of coupling strengths across the critical coupling strength.

Transfer entropy calculations focussed on the transient period in which those networks with above-critical coupling achieved synchronisation. This provides insight into the information dynamics during the distributed computation by the network of whether it will synchronise and what the synchronised phase will be. Calculations were made for every connected pair in the network, and produced average TEs for each pair, as well as local TE values (see Sect. 4.2.5) for every time step for each pair. The calculations used the time series of *relative* phases and phase differentials computed by the Kuramoto updates, rather than the absolute phases of each oscillator. This choice was made for similar reasons to that of relative speeds etc. for swarms (as discussed in Sect. 5.5):

1. The state update information is contained directly in the *relative* phases and phase differentials rather than proxied in the *absolute* phases.
2. This allows accumulation of observations in the PDFs of interactions which are dynamically equivalent despite occurring at different *absolute* phases.

TE was computed using box-kernel estimation [150, 298].

What did transfer entropy reveal about synchronisation? Ceguerra et al. revealed several new insights into the dynamics of synchronisation with TE here [54]. While these insights have only been observed on these locally connected networks, it is expected that they will generalise to other complex network structures.

First, Ceguerra et al. examined the dynamics of local transfer entropy (averaged across all nodes in the network at each time step) as the synchronisation process unfolded. The local TE was observed to take large values initially, but then drop away to zero by the time the network synchronised (i.e. when the distributed computation had finished because the system reached an attractor). This result was as expected, since once coherence is achieved the nodes' behaviour is predictable from their own pasts, and they execute information storage rather than transfer dynamics. More interestingly, however, was that:

Key Result 9: *The transfer entropy dropped to zero significantly earlier than the order parameter indicated that synchronisation had been achieved.*

This is an important result, indicating that the distributed computation of what will be the synchronised state was completed much earlier than synchronisation is observable using conventional application-specific measures.

Next, Ceguerra et al. observed that the average transfer entropy (over time and all node pairs) increased with the coupling strength between the nodes, as the system is moved from an incoherent to synchronised state (in the parameter space of the coupling strength). This is intuitive to a large degree, but is not always the case in other systems, e.g. see [11]. Even in this case, the intuition is unclear, because increased coupling strength meant shorter transient time for transfer to occur; yet, the result indicates larger transfer taking place in less time as the coupling strength increases.

Finally, Ceguerra et al. examine the relationship between the transfer entropy to or from each node, and the position of the node within the network structure. This revealed an interesting hierarchy in the network, including a computational core, with large incoming and outgoing transfer, and a communication shell, exhibiting large outgoing but low incoming transfer. Also:

Key Result 10: *Strong correlations were observed between node degree and outgoing transfer entropy*

This is perhaps because the larger diversity of inputs for high degree nodes provides more novel information to transfer to other nodes. Seemingly conflicting results were found using a related measure under different (Gibbs) dynamics [276], possibly explained in that these dynamics seem to constrain the diversity in behaviour of high-degree nodes. Further work is required to fully explore the relationship of TE to degree.

5.7 Summary

In this chapter, we have reviewed applications of the transfer entropy to several classic complex systems, including cellular automata, the Ising model, random Boolean networks, small-world networks, swarming models and synchronisation processes. These applications of TE are important because of the central position of these models in complex systems science and their relation to many other real-world systems, as well as previous conjectures about the nature of information transfer in their dynamics. In each case, *TE has brought us new understanding of the role of information transfer in the intrinsic computation in these processes*, for example the

fundamental result that gliders in CAs are indeed the dominant information transfer entities in those systems. As flagged in Key Idea 28, we have seen that using transfer entropy, even in these simple systems, requires some subtlety and thought into which information channels to measure and how to approach such measurement (e.g. whether any pre-processing is required, and which estimator to use). For example, we measured TE on relative variables in Sect. 5.5 and Sect. 5.6, used the dynamics of local transfer entropy to understand space–time dynamics in Sect. 5.1 and Sect. 5.5, and only gained full understanding of the computational behaviour of a system when combining TE with related quantities such as information storage in Sect. 5.1, Sect. 5.3 and Sect. 5.4.

These applications provide a solid theoretical grounding for the next chapters, which explore the application of TE to real-world data (or at least more detailed models), including in finance and economics in Chap. 6 and neuroscience and other application domains in Chap. 7.

Chapter 6
Information Transfer in Financial Markets

In 1900 a young French mathematician named Louis Bachelier published his PhD thesis entitled *Théorie de la spéculation* (The Theory of Speculation) [12, 65], and in many respects this remarkable work was more than half a century before its time. In it Bachelier described the statistical properties of the fluctuations in price movements in financial instruments and their derivatives. Perhaps the most fascinating aspect of Bachelier's thesis is that in it he develops a general theory of the statistical fluctuations that is used in a large variety of different applications [98] including ecology, chemical reactions and physics. This would later be called Brownian motion, predating Einstein's work [83] in this area by five years.

In this way the process of mathematising the fluctuations in financial instruments was begun. Perhaps the most important result in this area is the Black–Scholes equation [40] used for pricing certain financial derivatives. Myron Scholes would be awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel (Nobel Prize in Economics) in 1997 for his work in this area and would later play a more direct role in the financial markets by co-founding the company *Long-Term Capital Management* (LTCM). LTCM was a hedge fund founded in 1993 that was extremely successful in its early years, but after the 1997 Asian financial crisis and the 1998 Russian financial crisis, LTCM would lose around US $4.6 billion [203].

This loss is striking for a number of reasons. First, LTCM had some of the best *mathematical technology* available at the time in order to understand price movements: two Nobel prize winners were on the board (R. Merton and M. Scholes), both of whom had won the prize for their work on estimating prices in stochastic markets. Second, they had some excellent practical market experience in a former vice chairman of Salomon Brothers (J. Meriwether) and a vice chairman of the US Federal Reserve (D. Mullins Jr.). Finally, they were very well capitalised with over US $1 billion in initial funding in 1994 and by the time they collapsed had US $4.6 billion in equity.

So what went wrong? An intriguing perspective is provided by the *sociology of arbitrage* [209]. As circumstances change it is possible for a (comparatively) unrelated event to cause strong correlations between prices that were previously

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_6

uncorrelated. For example the Russian financial crisis may have been responsible for unwanted price correlations that were the onset of LTCM's trouble [209]:

> Crucially, correlations between the different components of LTCM's portfolio leapt upwards from their typical level of 0.1 or less to around 0.7. Suddenly, a whole range of positionsd—hedged, and with little or nothing in common at the level of economic fundamentals—started to incur losses virtually across the board. LTCM's losses were stunning in their size and rapidity: in August 1998 [the month of the Russian financial crisis], it lost 44 per cent of its capital.

These correlations are critical because an *uncorrelated* portfolio of stocks has a lower risk profile than a correlated portfolio [249]

So in very nearly 100 years of research, from 1900 to 1998, we have moved from trying to understand the *independent* price fluctuations through Brownian motion to understanding *correlated* price fluctuations driven by sociological factors. In this chapter we will look at what is currently understood about the drivers of financial and economic systems through the use of transfer entropy.

6.1 Introduction to Financial Markets

Financial markets such as the New York Stock Exchange (NYSE) or the NASDAQ stock market are a cornerstone of modern financial economics. Stock exchanges in particular enable companies to raise capital in order to grow through the funding of new projects by providing a place where investors can buy a stake in a company through the purchasing of shares. As a reward for investing in a company that performs well an investor can earn a profit through dividends, a percentage of the companies profits that are divided amongst the shareholders, as well as through selling the shares at a higher price than the original price of purchasing the shares.

One of the most important measures of overall market performance is provided by market indices. These are aggregate measures of the overall market performance of a subset of the equities traded on that market. For example the Standard and Poors 500 (S&P 500) is the weighted average of the 500 largest (by market capitalisation), publicly traded equities on the NASDAQ and the New York Stock Exchange. An alternative index for the US stock market is the Dow Jones Industrial Average (DJIA), an adjusted average of 30 publicly traded equities. A market index is a summary statistic of the performance of the equity market that has an in-built bias in terms of which equities are included and how they are weighted. From this point of view the NASDAQ Composite index (an index of technology stocks traded in the United States) summarises the technology market's performance, the DJIA summarises (approximately) the US manufacturing base and the S&P 500 (approximately) summarises the most highly capitalised equities. Other indices around the world reflect different biases and weightings on their respective markets such as the Frankfurt Stock Exchange (DAX), the London Stock Exchange (FTSE 100) and the Australian share market (All Ordinaries Index, AO).

The importance of these indices has extended beyond the simple summary of equity values. It is now common to cite the performance of a market index as a measure of the economic performance of a market sector or even the country as a whole. This extension of financial markets to national economic performance is important. A financial market is a mechanism by which companies can raise financial capital to fund their operations by selling off a portion of the ownership of their company in the form of shares. In turn the investor can expect to receive dividends from the company based on the profit performance of the company they own a part of. The investor can also make a profit by selling their shares in a company at a later date if the price of their shares has increased. These two ideas are related: a strongly performing company may be able to increase its dividends, and higher than expected dividends can push up the price of the shares.

So a strong performance in the operation of the companies that make up a financial market index may result in an increase in the value of the index, but there are many other factors that also influence the price of shares that are not directly related to the underlying performance of the company. The terrorist attacks on the World Trade Center and the US Pentagon on September 11, 2001 are a striking example: nothing changed regarding the underlying performance of the companies that were part of any index, certainly not in the days immediately following the attacks. Despite this disconnect between the attacks and economic performance, the DJIA initially dropped 684.81 points (7.14%), the largest single-day decline in its history until the 2008 Global Financial Crisis, and it took 40 days to recover from this fall [79]. The underlying performance of the companies that make up the economy did not decrease by more than 7% and then rebound over the following 40 days [57], but the investors were responding at least in part to what they expected would happen in the future, and the future looked very unsettled if not downright terrifying. From this point of view, sometimes markets move as a reflection of the underlying performance of the economy and sometimes it is due to effects that are not part of our nominal economic expectations but have a psychological impact on market expectations nonetheless (See Keynes' beauty pageant (Sect. 6.4.1).

However, the changes in prices of equities show some unusual behaviours that have made the study of their statistics a non-trivial matter. So in practice what is most likely being observed in the price variations is the very rapid diffusion of both relevant and irrelevant information through a financial market and its influence on how traders perceive the future value of individual equities. In this sense transfer entropy measures a combination of both *market sentiment* and *market fundamentals*. For an informative introduction to these two ideas see for example [15] and references therein.

So the share markets reflect to some extent the expectations of the market's performance and the economy as a whole, but the world is a non-linear place and expectations are often misguided. While markets themselves are exceptionally hard to predict with any level of assurance, there may be very broad drivers in the wider economy that can guide our expectations of what might happen in the financial markets. With this in mind there is considerable interest in finding out what the underlying drivers of our markets and economies really are: does unemployment drive the

GDP or is it the cost of imported goods as a function of the foreign exchange rate? So an interesting area to explore is the relationship between equities and indices, indices and indices, and indices and the economy as a whole in order to understand the extent to which changes in one financial or economic measure act as a precursor or driver to changes in the other. The following sections highlight some of the key findings in the recent literature where transfer entropy has been used in order to explain which aspects of finance or economics is driving another.

6.2 Information Theory Applied to Financial Markets

Information theory as applied to economics and finance has a history almost as long as that of information theory itself. In 1956, just eight years after Shannon's seminal 1948 article *A Mathematical Theory of Communication*, J. L. Kelly Jr. used information theory to prove what the optimal gambling strategy should be in many useful circumstances [155]; he also speculated on the possibility of using his result as an investment strategy for share traders. With this in mind we review some of the more interesting ways in which information theory has been applied in economics and finance.

6.2.1 Entropy and Economic Diversity: an Early Ecology of Economics

In 1975, Hackbart et al. put forward an interesting proposal [122]: Can you use entropy to measure the changing diversity of an economic system? The thought appears to have been inspired by the use of entropy as a measure of diversity in ecological studies of bio-diversity, see [148] for an overview of the most common measures used and their relationships. This seems to be remarkably prescient of Hackbart and colleagues in light of recent calls for an *ecology of economics* [219, 125], and particularly of banking systems in light of the 2007-2008 financial crises. This leads to an interesting interdisciplinary question:

Key Idea 29: *The notion of ecological diversity, as measured by entropy and its generalisations, can help us understand the interconnectedness, stability and sustainability of our modern financial systems.*

6.2.2 Maximum Entropy: Maximum Diversity?

If entropy is a useful measure of diversity, be the system ecological or economic, then is it possible to understand our ecologies or economies as maximisers of diversity via the maximisation of the system's entropy? The question arises because, returning to the physical notion of Brownian motion, maximising the entropy appears to be what some physical systems *do*. This includes the suspended particles considered by Einstein in 1905 that led to his work on Brownian motion and a vast array of other systems studies since then. E.T. Jaynes notably addressed this by showing that it is possible to reconstruct much of modern physics by maximising the entropy of an appropriately described (i.e. constrained) system [145, 146] without needing to know anything about the microscopic interactions of the elements of which the system is composed. This is called the MaxEnt method, and it provides a very straightforward way in which to construct probabilistic models of systems with otherwise opaque internal processes.

This has been an incredibly powerful heuristic in physics, and just as Brownian motion was used to describe both the physics of particle movements and the fluctuations in financial prices, it then leaves open an interesting question: To what extent can the MaxEnt principle be applied to economics and finance? Some progress has been made in this area, particularly in the direction of the micro-economics of gametheory. Several recent studies [364, 133, 134] have shown that maximising the entropy of each player's choice probabilities results in a generalised form of the classical Nash equilibrium called the quantal response equilibrium (QRE). The QRE has been empirically investigated as a model of bounded rationality in decision-making experiments [109, 368].

Key Idea 30: *Jaynes' MaxEnt principle can be used to model the decisions of economic agents in micro-economics.*

6.2.3 Mutual Information: Phase Transitions and Market Crashes

In keeping with the shared intellectual history of financial markets and physics, many researchers had noted that there is an analogy between market crashes in finance and *phase transitions* in physics (see Sect. 3.3). Some earlier studies observed that the significant property to a market crash was the level by which the whole market suddenly dropped as measured by a market index. However, the key to phase transitions in physics lies in the changes in the correlations between the interacting elements of the system: as a phase transition approaches, correlations grow stronger across the whole system. An early empirical study [160] into this analogy for financial markets looked at the S&P 500 index around the Black Monday crash of

1987 and the correlations between equities. It was found that the financial market behaves a lot like a physical system as it approaches a critical point via the variation in a control parameter. This is an explicit example of the non-stationary aspect of financial markets: there appears to be an underlying parameter of the system that varies over time, changing the shape of the probability distribution of price changes. See Sect. 2.3.6 for a discussion of non-stationarity.

This gives strong support to the notion that a market crash is in some sense like a phase transition in physics. However the drop in a market index does not tell us anything about the underlying interactions, and correlations can rise and fall across a range of values without necessarily going through a phase transition. The mutual information of a system is known to peak precisely at a phase transition in a multitude of different systems, from spin systems in physics [218] to more general systems beyond physics [353]. So a recent study looked at whether the mutual information across a large portion of a financial market peaks during a crash [132], thereby providing an important piece of the puzzle in forming the analogy connecting market crashes with the physics of phase transitions. The data was taken from a 13-year period spanning several key crashes including the 1997 Asian financial crisis, the 1998 Russian crisis, the dot-com collapse and the beginning of the 2009 market collapse. All of these events showed clear changes in their information measures at the known critical points.

Key Idea 31: *Information theory can be used to analyse the critical phenomena of financial markets, such as market crashes, just as it can be used in other complex systems.*

6.3 Information Transferred from One Market Index to Another

Perhaps the most natural question that we might ask of the relationship between two financial indices is: *To what extent does one index drive the behaviour of another?* For example we might use the NASDAQ index as a proxy for the technology industries and the DJIA as a proxy for the industrial and manufacturing industries. Then if we were to use transfer entropy to study in which direction the information flowed from one index to the other we would get an indication of the extent to which the expectations of one industry's performance influences the other. In practice this is an imperfect measure: the indices only measure aggregate price movements and aggregate expectations, not the underlying economic fundamentals, so the buying and selling patterns of the traders fluctuate more rapidly than any real economic variation in the businesses they buy equity in. This implies that, whatever connection exists between the indices, it is not going to be on an economic time scale of months or years, it will be on the time scale at which information diffuses through the marketplace, and given the speed of modern communication, what people are

saying now will be reflected in market prices within hours if not minutes. It is also important to note that information will typically flow in both directions, i.e. for two time series X and Y: $\mathbf{T}_{X\to Y} > 0$ and $\mathbf{T}_{Y\to X} > 0$, so more often than not we are interested in the *net information flow* (NIF): $\mathbf{T}_{X\to Y} - \mathbf{T}_{Y\to X}$ as an indicator of which time series has the stronger influence over the other.

Marschinski and Kantz, one of the first research teams to tackle the question of index-to-index analysis using transfer entropy, considered the relationship between the DJIA and the DAX between May 2000 and June 2001 (63,867 data points) [216]. In this work the DJIA and the DAX are analysed using the first application of the transfer entropy by an approximation called the *effective transfer entropy* (ET) (see Sect. 4.5.2) and the *relative explanation added* (REA), both information-based measures of causal relationships. The ET addresses a key issue with estimating TE from real data: because the TE is always non-negative, it has a tendency to over-estimate the TE signal, i.e. it is a biased estimator of the real TE (see Sect. 3.2.2 for a discussion of biases in entropy estimation, and Sect. 4.5.2 for ET). In order to reduce this bias the TE can be calculated for the original data and then recalculated using the same data but shuffled to remove any relationship between the two data sets. The shuffled data will have a TE which is typically greater than zero (due to random variations resulting in coincidental shared information between the data sets) that can be subtracted directly from the measured TE to give the effective transfer entropy:

$$\mathbf{T}^{\text{eff}}_{X\to Y} = \mathbf{T}_{X\to Y} - \mathbf{T}^{\text{shuffled}}_{X\to Y}. \tag{6.1}$$

The second new measure introduced in this work is the REA (for notational consistency labelled $\mathbf{R}$), a measure of how much of the total information flow $H_Y(m)$ in series Y from the last m time periods is explained by the effective transfer entropy from X to Y based on the last n time periods in X:

$$\mathbf{R}_{X\to Y}(m,n) = \frac{\mathbf{T}_{X\to Y}(m,n)}{\mathbf{H}_Y(m)} \in [0,1]. \tag{6.2}$$

$\mathbf{R}_{X\to Y}(m,n)$ can be interpreted as the percentage of the total entropy transfer in Y that is explained by the effective transfer of entropy from X.

The key finding in this work is that, at the time scale of less than a minute, the DJIA $\to$ DAX has approximately three times the REA signal (average 1.25%) compared with the information flow as measured by the REA in the opposite direction DAX $\to$ DJIA (average 0.42%). So the net effective transfer of entropy was in the direction DJIA $\to$ DAX (average 0.83%).

In order to understand this result for market indices $X(t)$ and $Y(t)$, they compared it with an equivalent linear autoregressively coupled system of the following form:

$$x(t) = r(t) + \varepsilon y(t-1). \tag{6.3}$$

where $r(t)$ and $y(t)$ are random Gaussian noise with zero mean and unit standard deviation. Note that ε controls the strength of the coupling between $x(t)$ and $y(t-1)$. Using this model they found the ε value that most accurately reflects the REA observed between the two indices: for DAX $\rightarrow$ DJIA, $\varepsilon \simeq 0.1$ and DJIA $\rightarrow$ DAX, $\varepsilon \simeq 0.2$. These coupling strengths translated into an ability to forecast the direction in which one time series was going to shift from another approximately 56.5% of the time, despite the author's caution that the actual coupling between these time series is a non-linear one.

A second study [281] used a similar approach but applied it to the log price returns of the Indian stock market using the Nifty index and the US$/Indian rupee exchange rate (FOREX) as the two time series. Unlike the DAX and the DJIA, this is a comparison between two financial markets using one index within the country (the Nifty) and one on the economic border of the country, as the US$/rupee exchange rate might usefully be thought of as the border between the internal markets of India and the financial world denominated in US$ (this simplified world view excludes speculative traders who have no fundamental economic interest in India). Much like the previous study, the REA and the NIF were used, and a third measure was introduced, called the normalised directionality index:

$$d(X,Y) = \frac{\mathbf{T}_{X\rightarrow Y} - \mathbf{T}_{Y\rightarrow X}}{\mathbf{T}_{X\rightarrow Y} + \mathbf{T}_{Y\rightarrow X}} \in [-1,1]. \tag{6.4}$$

The term "normalised" is a misnomer in the statistical sense as $d(X,Y)$ is not bounded between 0 and 1 as a (normalised) probability distribution would be. Instead it more closely resembles a measure of divergence or market leverage as it is maximised when one of the TE values is zero and minimised when they are equal. However, it does regularise the TE measure such that $d(X,Y)$ will always lie between -1 and 1, a useful approach when we want to compare measures across different partitions or different systems.

Over the time period considered (November 1997 to March 2007) a small net information flow was detected where the dominant direction was from the Nifty to the FOREX market. At times the $d(X,Y)$ measured for Nifty $\rightarrow$ FOREX reached values of 1 as the flow in the FOREX $\rightarrow$ Nifty direction dropped to zero. But these weak signals may be an indicator of the time frame over which information diffuses through financial markets. Over the period of a whole day (the finest resolution of the data set used) price data has likely already integrated most of the information that has passed from one market to another, and hence the market prices have equilibrated with respect to this information. This is peculiar to modern financial markets: the speed with which we can communicate with each other as well as the speed with which we can trade on the markets has made day-close price signals less informative than they have been in the past.

These two data sets had very different resolution scales: the DAX and DJIA data set recorded every single increment on a second-by-second basis, whereas the Indian Nifty and FOREX data were recorded at daily intervals. This enabled the DAX/DJIA study to detect significant signals at the sub-minute level, suggesting

that information encoded in equity prices diffuses through the market incredibly quickly. This information dissipated quickly and markets absorbed it, or equilibrated, very quickly. This was also true for the Nifty/FOREX data set: the signals were very weak at the daily level making it difficult to draw strong conclusions. However, what was made clear is that, if data is of a fine enough resolution, then conclusions regarding the amount and direction of market influence could be derived from financial data, opening up the possibilities of new ways in which both financial and economic signals could be analysed.

6.4 From Indices to Equities and from Equities to Indices

Market indices such as the S&P 500, the DJIA or the FTSE 100 act as indicators of overall market conditions, sometimes with a deliberate bias reflected in the composition of a particular index. If a market index increases over a single day's trading then the equities that make up the index have, on weighted average, done well for that day. Likewise if the index has decreased over the day's trading then the indexed equities have performed poorly. Analysts and economists can also refer to particular indices in order to indicate how the market (or market sector depending on the index) is performing without too great a concern for the actual composition of the index; for example an analyst might suggest the industrial sector is doing relatively well by comparing the increase in the DJIA over the last 3 months with the relatively sluggish performance in the technology sector over the same period as indicated by the NASDAQ Composite index. Neither of these indices represent the performance of the entire industrial or technology sector, but they do provide reasonable approximations of these market sectors.

These indices raise some very interesting questions, both theoretical and practical, about the nature of market dynamics and how people choose to buy and sell equities. To see this, imagine you are a trader and you want to know which equity to buy. You are confronted with choosing between many thousands of different equities, each of which represents a company with unique product(s), performance history, strategy, CEO management style, opportunities and risks associated with it. Potentially you could go through each company's annual reports, study its market position and strategy, look at what the CEO plans to do and whether or not the CEO's experience is adequate, and base your buying decision on the equity that you believe has the greatest financial opportunity given today's price for the given equity. This is called *fundamental analysis*, and it can be time consuming and difficult. On the other hand you could simply look at whether or not the market (or a particular market sector) is doing well by referring to an index and use that as a part of the analysis that goes into making a decision, so-called *technical analysis*.

This raises an interesting question:[1]

[1] Indeed, this question is evocative of the exploration of information flow from lower to higher system levels, and viceversa, in an exploration of top-down causality by Walker et al. [342].

Key Idea 32: *In which direction does the net information in markets flow, from the equity to the index or from the index to the equity?*

This is theoretically interesting because there are multiple possible models for random fluctuations that are well studied in the physics of particle systems. Each individual particle's behaviour fluctuates depending on how it is connected to the other particles in the system. If none of the particles are connected to each other, then if one particle changes its state then this does not influence any of its neighbours, and so they do not interact with one another. Such systems behave in a very stable fashion, and this might be a theoretical approximation to making your equity purchase decisions based on fundamental analysis. If on the other hand particles are linked to one another, say in a lattice-like fashion (for example imagine a two-dimensional grid connecting many thousands of particles together, similar to the Ising model in Chap. 5 (Sect. 5.2)), then changes in one particle's state influence their local neighbourhood through the connections between them.

Another possible arrangement is where each particle is actually connected to every other particle in the system in such a way that each particle is influenced by the average state of all of the particles in the system; this is referred to as a "mean field" model of particle interactions, and it is analogous to making your equity purchase decisions based on a market index (i.e. a weighted mean value of the market performance). Each of these three models can be a very powerful approximation to real physical systems, and there is considerable research into understanding the relationship between such models and financial and economic systems in the nascent field of *econophysics* [212].

6.4.1 Economics of Beauty Pageants

So with these ideas in mind consider that John Maynard Keynes introduced the idea of financial markets as a beauty pageant in his book *The General Theory of Employment, Interest, and Money* [1936] in order to explain price fluctuations in financial markets. In Chapter 12 he writes [156]:

> It is not a case of choosing those [photos of pageant contestants] that, to the best of one's judgement, are really the prettiest, nor even those that average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees.

Such a view of the financial markets opens us up to the possibility that a great deal of buying and selling (i.e. market price formation) is based on what each trader believes the average buyer or seller thinks *everyone else thinks* an equity is worth, and that everyone else is doing the same thing. However it seems most unlikely that *all* traders are following this strategy; only an unknown and perhaps unknowable

portion of market traders follow such a strategy, with the remainder of the traders basing their decisions on the underlying financial and economic fundamentals of the businesses they are trading in.

So could we, in principle, discover the degree to which traders base their decisions on average market sentiment, i.e. the ratio of fundamentalists to technical traders (opinion followers)? The answer is implicitly yes, and Kwon and Oh have done so across nine different market indices covering developing and developed markets in Europe, the UK, North America and the Asia–Pacific region [171]. Across all nine indices there was an unequivocal net information flow from the index to the individual equities with a ratio of $\mathbf{T}_{\text{idx}\rightarrow\text{eq}}$:$\mathbf{T}_{\text{eq}\rightarrow\text{idx}}$ ranging from approximately 3:2 to nearly 3:1. Curiously, while the effective transfer entropy was not calculated, the shuffled transfer entropies were plotted, showing that the shuffled values for $\mathbf{T}_{\text{eq}\rightarrow\text{idx}}$ were very similar to the unshuffled values, implying that there was little to no transfer of information from the individual equities to the index across all indices. The strongest $\mathbf{T}_{\text{idx}\rightarrow\text{eq}}$ signals appeared in the mature markets (S&P 500, FTSE (UK), NASDAQ, Canada, Italy and Australia), and they were weaker but still significant in developing markets (Thailand, Korea and China).[2] Such studies provide considerable empirical insight into market dynamics that will allow researchers to distinguish between the different candidate models that can be used as alternatives to previous, weaker models that need to be revised in light of recent market catastrophes.

6.5 The Internal Economy and Its Place in the Global Economy

It has been said many times and in many different contexts that we live in a world of unprecedented interconnectivity and the long-term effects of this interconnectivity are not yet well understood. In the previous section the different interconnectivities between individual equities and market indices were considered, showing a degree of synergistic interaction between equities based on market signals communicated through market indices.

Given the broad range of measures, both across countries and within countries, which have been constructed using transfer entropy, a final but more complex question is: *To what extent can transfer entropy be used to measure economic and financial signals both between and within countries?* This is an important question and it is similar to many of the most difficult questions research in complex systems addresses [159]. In this study transfer entropy is used to look at economic time series both within (using five macro-economic variables) and between 15 countries in order to establish an international economic influence network.

[2] In a previous paper, *Information flow between composite stock index and individual stocks* [2008], these same researchers looked at the net information flow from the DJIA, the S&P 500 to 125 individual equities between June 1983 and May 2007. They showed that the net information flow is from the index to the equities.

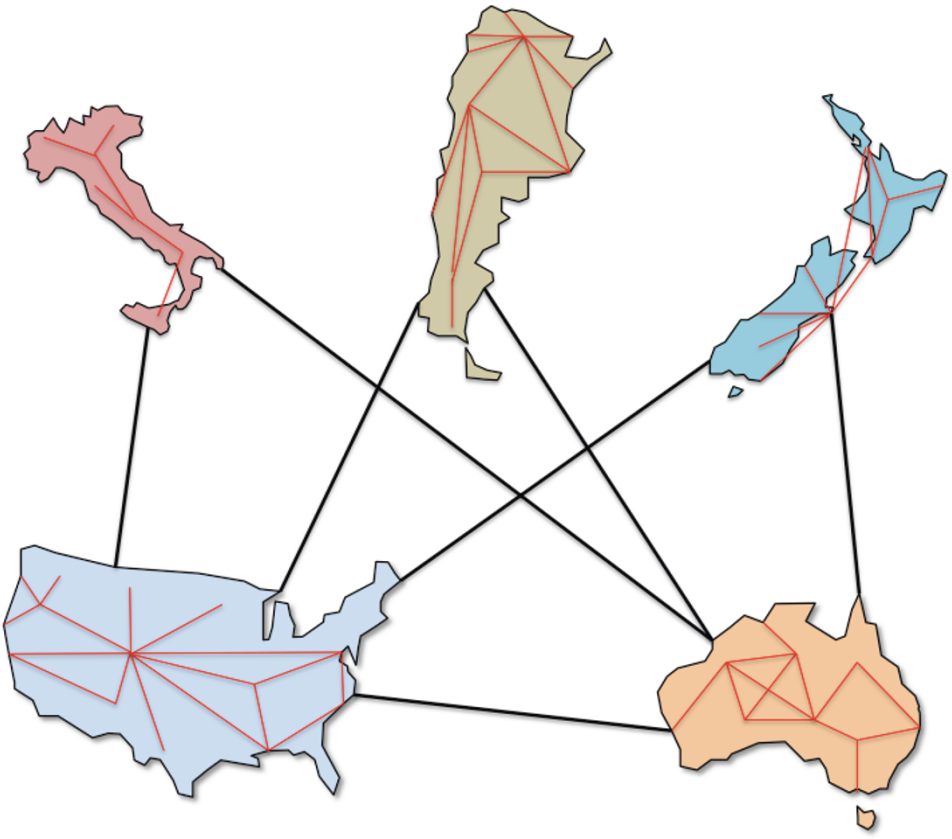


Fig. 6.1 Each country is connected to a number of other countries through a global network of economic relationships. Internally a country is governed by social, political, economic and geological constraints and relationships such as transport networks, natural resources, manufacturing centres as well as less obvious networks of social and political influence. These in turn are reciprocally coupled to the internal dynamics of other countries through trade, foreign exchange markets, political relationships and geographical considerations. Understanding how these factors influence one another, in particular the strength and direction of the connections, is of key importance for our understanding of how stable and sustainable our socio-economic systems are

One way in which we can understand the interconnectivities both within and between countries is to look at variables that reflect a country's internal economic changes and those variables that reflect a country's external economic and financial changes, i.e. changes in economic indicators between different countries. Kim et al. [159] took data from 18 different countries, encompassing Europe, North and South America, Asia and Africa; this included most of the G20 countries plus Spain and Portugal. They then collected five micro-economic indicators of economic performance, i.e.consumer price index [CPI], industrial production index [IPI], exchange rate [XR], stock market index [SMI] and trade balance [TB], in order to build a net-

work of relationships between a country's different economic time-series data and between different countries economic time-series data.

The result is an intriguing examination of the network of macro-economic relationships that make up our ever more interconnected economic world (Fig. 6.1). What the authors were able to show is:

Key Idea 33: *Western countries are globally the most influential, and Japan has become less influential following the Asian financial crisis in 1997.*

The Asian financial crisis, as a phase transition, has previously been studied using information theory in a different form [132], but what this work contributes beyond this earlier work is that the Asian crisis had a significant, long-term impact on the macro-economic relationships around the world, not just on the short-term financial market dynamics. This significantly extends the general reach of these information-based techniques to include short-term financial fluctuations in the order of days or weeks to extensive changes in economic performance over the range of years or even decades.

Kim et al. [159] also showed that influence is transferred more significantly between countries in Europe than in Asia or the Americas. This indicates that Europe is a much more tightly coupled economic system than other regions of the world. For example they showed that the German stock market index is a net information receiver from countries such as France, Italy and Portugal while on the other hand Portugal's stock index is a net information source for countries such as Germany and Italy. Similarly, Italy receives considerable information from France, Portugal, Spain and the UK. The significance of understanding these information transfers within Europe is highlighted by the recent economic turmoil in the region; Portugal and Spain have gone through significant economic difficulties as reflected in their broad macro-economic indictors such as unemployment rates, stock market performance and gross domestic product. Perhaps the most important result that is emerging from this collective work is:

Key Idea 34: *Understanding both the strength and the direction of macro-economic indicators provides an important insight into the knock-on effects that other countries feel as a result of a country's internal economic distress.*

The financial and economic relationships that make up a significant component of our individual wellbeing, whether it is our retirement funds, our mortgages or our holiday savings, are a vast interconnected network that influences nearly every facet of our lives. In this chapter we have shown that transfer entropy as a measure of the inter-relatedness between financial indicators has the potential to elucidate the everyday dynamics of the world's economies, from the moment-by-moment trades of the stock markets to the connectivities of modern global networks of trade and

economic health and prosperity. What this type of micro-to-macro economic analysis will tell us in the future is uncertain, but the promise it holds for analysis and diagnosis is extremely great and hints at the potential "model-free" tool that will help frame the next evolution in economic theory, analysis and modelling.

Chapter 7
Miscellaneous Applications of Transfer Entropy

The previous chapters have outlined the use of transfer entropy in some of its major application domains: providing insights into canonical complex systems and financial markets. Here, we complete the book by providing a high-level view of the wide-ranging use of transfer entropy in a suite of other fields. Our survey here highlights the importance of the measure and its ability to provide new insights about information flow across an impressive breadth of application domains.

In the following sections, we describe applications of transfer entropy in physiological data, inferring effective networks from multivariate time-series data, computational neuroscience data sets, biochemical networks, embodied cognitive systems and social media.

7.1 Information Transfer in Physiological Data

We begin here by considering large-scale physiology as one of the earliest application areas of transfer entropy. Indeed, Schreiber included an application to heart–breath rate interaction in the original presentation of the transfer entropy [298]. TE analysis is attractive for physiological data, due to the increasing level of automation in this domain, as well as the apparent complexity—coupled with uncertainty of causes—of the non-linear interactions in such data sets. In this section, we first review Schreiber's application of TE to heart–breath rate data, as well as subsequent investigations of the same data set. These analyses serve to highlight that:

Key Idea 35: *TE can give quite complex answers, even for apparently simple questions, and remind us of the care required in selection of estimators and parameters in order to achieve robust and reliable results.*

Afterwards, we also review the application of TE to other physiological data sets.

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_7

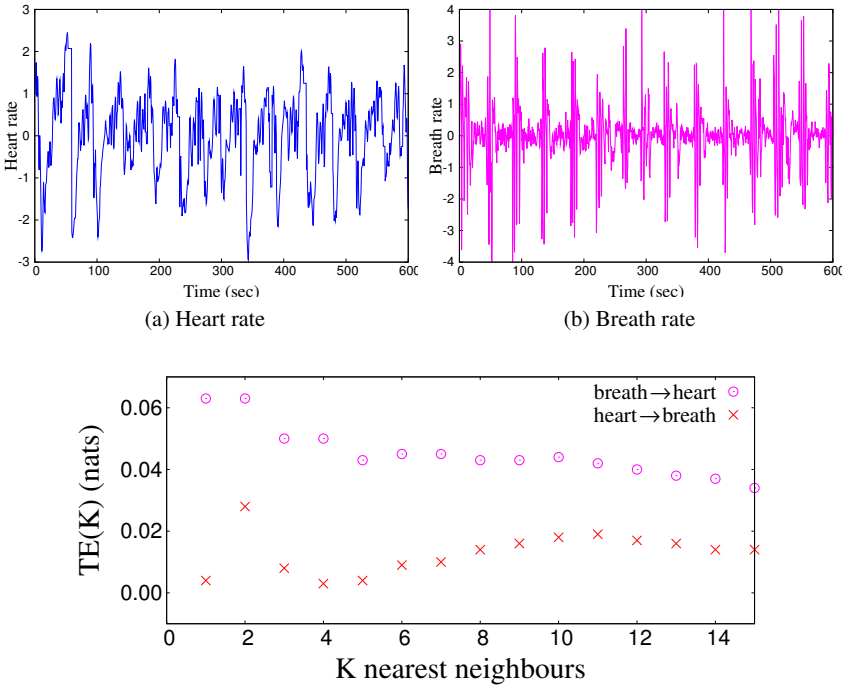
As stated above, Schreiber [298] included a TE analysis of heart and breath rate time series in the original presentation of the transfer entropy. The data were recorded from a sleep apnoea patient,[1] made available via the Santa Fe Institute time series contest held in 1991 [284] (shown in Fig. 7.1). Schreiber normalised each time series to zero mean and unit variance, then computed TE using box-kernel estimation (see Sect. 3.4.1.4) with history lengths (or embedding dimensions) $k = l = 1$, while varying the common kernel width applied to the two normalised time series over a range of values. That study found TE in both directions between heart and breath rate, indicating a complex interaction between heart and breath. With that said, there was greater transfer from heart to breath (over a wide range of kernel widths), which seemed to be in alignment with the observation of apnoea events occurring when the heart rate crosses some threshold.

Kaiser and Schreiber [149] soon revisited this analysis, suggesting that incorporating a common kernel bandwidth for both normalised series was unsatisfactory "since the two signals are of different physical nature and cannot be easily compared". They re-analysed the data, again with $k = 1$, by applying pairs of different kernel widths for the two time series, to examine the trends in TE as a function of these different scales. They interpreted their results to suggest that, since the TE values do not converge, and since different directions (heart $\rightarrow$ breath and breath $\rightarrow$ heart) dominate in different bandwidth regimes, one could not conclude on which signal was the driver and which the responder. In Schreiber's words from the original study: "Reducing the analysis to the identification of a 'drive' and a 'response' may not be useful and could even be misleading" [298].

Later analyses continue to suggest a complex two-way interaction here. Ancona et al. [3] used a related non-linear Granger causality measurement (using fixed-width radial basis functions) to search for an appropriate embedding dimension (or history length) for this data set, settling on $k = 5$, which also corresponded to the periodic respiratory rhythm. Ancona et al. then demonstrated consistently larger transfer from heart $\rightarrow$ breath, though they did find strong transfer in the opposite direction as well. An important aspect of their findings is the use of a proper embedding of the data (which rules out the mistaking of storage as transfer—see Sect. 4.2.2 and the CA example in Sect. 5.1).

In addition to proper embedding of the data, we note that it would be useful to add bias correction to this analysis, and remove the arbitrariness of selecting kernel widths. As such, we describe an application of the KSG TE estimator [168, 110] (see Sect. 4.3.1) to this data set. Since this estimator automatically scales the *kernel width* for each observation in the marginal dimensions (while keeping a fixed neighbour count in the full joint space), and is quite stable with respect to parameter selection, it may be more suited to addressing Kaiser and Schreiber's concerns [149] regarding comparing two signals of different natures. This demonstration is distributed with the open-source *Java Information Dynamics Toolkit* (JIDT) [183] (i.e. the Matlab/Octave script `runHeartBreathRateKraskov.m` in the `demos/octave/SchreiberTransferEntropyExamples` example of the distribu-

[1] Sleep apnoea is a disorder characterised by pauses and bursts of breathing during sleep—see the bursts and fluctuations of heart and breath rate in Fig. 7.1.



(a) Heart rate (b) Breath rate

(c) Transfer entropies

Fig. 7.1 Transfer entropy in heart and breath data, using a KSG estimator, as a function of the K nearest neighbours parameter of the estimator (using embedding lengths $k = 2$ and $l = 2$ as determined in the demonstration). Note the stability of the results for $K \geq 4$, in particular for the heart $\rightarrow$ breath measurement, which is well above the noise floor for statistical significance (not indicated on the figure; see Sect. 4.5.1 for details on how to compute this). For $K < 4$ the results are affected by under-sampling. Estimations are produced by the MATLAB/Octave script `runHeartBreathRateKraskov.m` in the `demos/octave/-SchreiberTransferEntropyExamples` example distributed with the Java Information Dynamics Toolkit [183]. Fig. 7.1a and (b) were first published in [182]

tion). The demonstration reveals more consistent TE measurements on the heart–breath data set with this estimator (with respect to parameter changes) than with the aforementioned kernel estimators – see Fig. 7.1. These measurements again find significant TE in both directions, albeit with a consistently stronger transfer in the breath $\rightarrow$ heart direction in contrast to Schreiber's original results with kernel estimation.

A clear conclusion from the above is that:

Key Result 11: *TE analysis is difficult to get right, and is best performed using estimators which are stable with respect to parameter changes (in particular the KSG estimator). One should take care with such parameters, as well as ensuring that data is embedded correctly.*

Another clear conclusion from the collection of analyses above is that we have a complex *two-way* interaction occurring between the heart and breath rate, which cannot be simplified to drive and response dynamics. In such situations—and even where we do have simple drive–response systems—more subtle approaches can be particularly revealing. For example, Lungarella et al. [205] study a different heart–breath sleep apnoea data set using transfer entropy applied to wavelet-transformed states. Their study again reports strong transfer in both directions, with heart $\rightarrow$ breath transfer being larger in most frequency bands involving bidirectional transfer, but with breath $\rightarrow$ heart transfer almost uniquely present at low frequencies. Other specific insights into the original heart–breath data set have been generated by using local TE-style approaches (see Sect. 4.2.5). Lizier [182, section 8.1] reported a preliminary study of local TE values at each time point in the time series, using kernel estimation with a single kernel width, and embedded history length $k = 4$. The time series of local TE values indicated significant bidirectional information exchanges coinciding with the apnoea events, with little transfer taking place in between them. Crucially, during typical apnoea events, the information exchange was started by a significant transfer from heart rate to breath rate; it was suggested that this precedence in time is more indicative of dominance in the dynamics than the relative average TE values. Williams and Beer [359] provide a more in-depth analysis, focussing on a *partial information decomposition* (see Sect. 3.2.3.1 and [358] for details) of the transfer entropy measurements (kernel estimation, with $k = 1$). Their key insight was to determine the transfer entropy in a partially localised manner: as a function of the target's past state. This revealed for example that transfer from heart to breath was largest when chest volume was low, and to a lesser extent when it was high, but was minimal when chest volume was near its mean. The results held over a wide range of kernel widths, and again provide far deeper insight into the dynamics here than average TE values.

The aforementioned studies on this data set demonstrate not only the breadth of ways that TE can be used to investigate a system, but also the complexity that it can reveal.

Of course, applications of TE in physiology have not been confined to this one data set. In particular, Faes and colleagues have utilised the TE to provide insights into a number of physiological processes [85, 86, 87]. For example, Faes et al. [87] investigated cerebrovascular and cardiovascular regulation in patients exhibiting orthostatic syncope a transient loss of consciousness and postural tone with spontaneous recovery. They sought to characterise the methods of such regulation from an information processing perspective, seeking to provide new insights and possibly early detection of syncope events.

Faes et al. analysed three sets of time-series measurements (300 samples, 1 per heart beat) per patient (10 patients), derived from ECG signals, recorded: (i) during a preceding lying position, (ii) just after a transition to a tilted position intended to evoke a syncope event and (iii) just before the syncope event occurs. Transfer entropy was measured (in addition to other measures of information dynamics) between heart period (HP) and systolic arterial pressure (SAP) to study cardiovascular regulation, and between cerebral blood flow velocity (CBFV) and mean arterial pressure (AP) to study cerebral regulation. A novel non-uniform embedding was used to represent the past history of the target variable (see details in [87]). Entropy estimation for the TE was performed using a uniform six-bin quantisation or discretisation, as well as a bias correction term (described in [265]). The TE results revealed that in the lead-up to a syncope event, TE from SAP to HP decreased significantly, while TE from AP to CBFV increased significantly. Faes et al. conclude that these insights identified deficiencies in the regulation mechanisms, and importantly characterised the impairment of the mechanism in a manner that differentiated between conflicting earlier hypotheses.

7.2 Effective Network Inference

A key objective of multivariate data analysis in many domains is to infer a *network* which underpins the observed activity levels between individual variables in the data. That is: *given only time series for each of a set of variables, can we describe a network which represents the relationships between these variables?* This line of inquiry is particularly popular in computational neuroscience (see [351] and chapters therein), but has also been addressed in other domains, including financial markets, gene regulatory networks and social media.

There are three fundamentally different notions of the type of network one may try to infer *functional*, *structural* and *effective* networks. *Functional network* inference constructs *undirected* networks using a measure of correlation between nodes to infer connectivity [95]; while this infers relationships between nodes with similar dynamics, it provides no explanation for how the relationship manifests. *Structural network* inference seeks to reveal the physical, directed (causal) connections in the system, though this is generally only possible via *interventional* techniques but not directly from large (observational) multivariate time-series sets alone [11, 261, 191, 60].[2] Structural networks also do not tell us about time or experimentally modulated changes in how the variables are interacting [95]. *Effective network* inference is something of a middle ground between these:[3]

[2] It can be done under certain circumstances from observational data, e.g. using the "back-door" approach [191, 11], though this typically requires a priori knowledge of where all other causal links are to the given node, defeating the purpose of general inference.

[3] "Model" here has several interpretations. Friston generally uses the term for a specific type of model (as in dynamic causal modelling) [96]; others use the term to refer simply to models with

Key Idea 36: *Effective network analysis examines* directed *(time-lagged) relationships between nodes from their time-series data, and seeks to infer the "minimal neuronal circuit model" which can replicate and indeed* explain *the time series of the nodes [311, 95].*

An effective network should reflect the underlying structural network, however it is not intended to give a unique solution to structural network inference from a time series, since it should be experimentally dependent and time-dependent (to capture the result of external modulations) [95].

Key Idea 37: *Transfer entropy has been recognised by the research community as a natural fit for effective connectivity inference, since it measures the directed relationship between nodes in terms of the predictivity (or explanation) added by the source node about the target.*

As we will describe in the following, transfer entropy has been used to a very large extent for effective network analysis in computational neuroscience, financial market analysis, gene regulatory networks, social media and multi-agent systems. In comparison with the related Granger causality, which has also often been used for this purpose, transfer entropy is of course model free and captures non-linear interactions.

7.2.1 Standard Pairwise TE Approach for Effective Network Inference

Early approaches to using transfer entropy for effective network inference used the following *basic* algorithm [139, 189, 73, 142, 32, 334, 296]:

1. Measure pairwise TE between all pairs of variables in the system;
2. Threshold the TE values to select connections for the network.

Standard approaches to using (pairwise) transfer entropy for effective connectivity analysis now however use approaches of *statistical significance testing* to determine whether links should exist. That is, the above basic approach is changed to [350, 336]:

1. Measure pairwise TE between all pairs of variables in the system;

---

directed connections; whilst we use the term to mean a model that is capable of producing the same computations.

2. For each source–target pair, generate the null distribution for TE (under a null hypothesis of no source–target temporal relationship, as described in Sect. 4.5.1), and obtain the p-value for measuring the observed TE under the null distribution;
3. Threshold these p-values to select connections for the network (i.e. selecting those with low p-values).

The generation of such p-values for TE was originally described in [335, 56] (as per Sect. 4.5.1), and first used for effective network analysis in this way in [337, 349, 187].[4] Since inaccuracies in TE estimation are unavoidable, the shift to statistical significance adds robustness to the approach and makes it more suitable for small data sets [337, 187]. The choice of p-value threshold is also more principled than a fixed TE threshold, since it can be meaningfully applied simultaneously to variables with very different statistical properties (which may impact their raw TE values), and allows a very specific statistical interpretation of the meaning of an inferred connection.

Correction for multiple comparisons using family-wise error rates (e.g. Bonferroni correction) or false discovery rates becomes particularly important when one is testing the statistical significance (with respect to a given p-value) of a large number of pairs in the potential network here [337, 187]. To further avoid false positives, the importance of proper embedding for the source and target variable was emphasised in [337], as well as measuring TE at the appropriate source–target delay [348, 337, 142]. Usefully, it has been established in this context that TE is somewhat robust to being measured at a smaller interaction delay [337] (with adequate embedding or when the source has memory), as well as for under-sampled data sets [187].

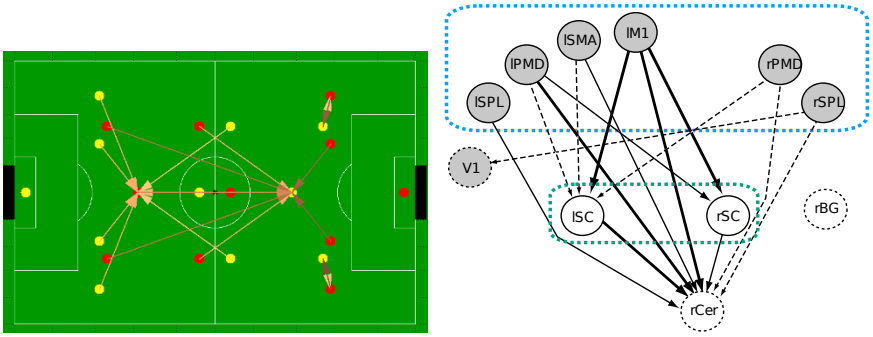
Sample networks output by these techniques are shown in Fig. 7.2 for interaction diagrams between players in robotic soccer and between brain regions from fMRI data.

Whilst effective network analysis is not strictly intended to replicate underlying structural networks, one generally expects results to closely match underlying structure when inferred over a large amount of data, sampling large ensembles of dynamics and experimental or input conditions. Under these conditions, receiver operating characteristic (ROC) curves are used to evaluate performance in terms of true positive versus false positive rate in comparison with an underlying structure, as a function of (p-value) threshold, e.g. see [142, 32, 201].

7.2.2 Addressing Redundancy and Synergy in the Data

It is widely acknowledged that the standard approach on its own is susceptible to inferring false positives $A \to C$ due to *cascade effects* where we actually have $A \to$

[4] Tung et al. [329] used a similar technique, but generated surrogates with the target variable perturbed, which is not recommended since it destroys the past-next state relationship for the target.



(a) Network of Robocup interactions (b) Inter-regional structure in fMRI data

Fig. 7.2 Sample effective network diagrams generated using TE-based algorithms. (a) A (simplified) network inferred from interactions between player movements in simulated football (Robocup) in [62]; the network reveals the central midfielders as information hubs. (b) An inter-regional effective network inferred from fMRI data recorded during a visuo-motor task in [187] (see definition of acronyms for regions in this paper); the network reveals a three-tier structure with a premotor-motor cortical sub-network (movement planning) at the top sending information to the superior colliculi (guiding eye position and attention) at the middle tier, and both then sending information onto the cerebellum (motor control) at the bottom. Fig. 7.2a is reprinted with kind permission from Springer Science+Business Media (© holder) from [62]: O. M. Cliff, J. T. Lizier, X. R. Wang, P. Wang, O. Obst, and M. Prokopenko, "Towards quantifying interaction networks in a football match", in S. Behnke, M. Veloso, A. Visser, and R. Xiong, editors, "RoboCup 2013: Robot World Cup XVII", volume 8371 of Lecture Notes in Computer Science, pages 1–12. Springer, Berlin/Heidelberg, 2014. Fig. 7.2b is after [187]

$B \to C$, or *common driver* effects where we actually have $B \to A, B \to C$. These false positives are due to **redundancies** between A and the true source B (see Sect. 4.2.3 and Sect. 3.2.3.1).

Early enhancements to the standard technique above to address these situations relied on heuristic post-processing of the inferred effective network. For example, Tung et al. [329] suggested the removal of the edge with smallest TE from an initially inferred triangle motif. Wibral et al. [349, 352, 348] introduced more sophisticated methods using the reconstructed source–target interaction delays on the edges of triangles in order to remove suspected cascade and common-driver effects. On a related note, Wibral, Vicente and co-authors [337, 349] also introduced a "shift test" to eliminate false positives due to instantaneous linear mixing.

Beyond heuristic techniques, it is known that multivariate conditioning approaches with the TE can be used to directly eliminate redundancies [195, 335]; e.g. a conditional TE $\mathbf{T}_{A \to C|B}$ would not make a false inference in the cascade effect scenario $A \to B \to C$ since the redundant information in A *and* B about C is conditioned out. Not only that, but such conditioning (or even the multivariate collective TE, see Eqn. 4.20 [187]) will additionally capture any **synergistic** effects on the target produced by the source in combination with the conditional variable(s) [195, 187]; e.g. a conditional TE $\mathbf{T}_{A \to C|B}$ would be able to make a correct inference

in the scenario $\{A, B\} \to C$ where $C = A$ XOR B (see Sect. 3.2.3.1), when pairwise measurements could not detect either source.

The importance of including such conditioning for inference of effective connections was emphasised in [332, 316], yet it is not immediately clear how this should appropriately be done for large data sets. For example, Quinn et al. [277] suggest conditioning on *all* of the remaining variables at once (i.e. using complete transfer entropies); however this may eliminate too many links if there is a large amount of redundancy in the network, and typically results in significant under-sampling for realistic data set lengths—even for multivariate Gaussian data [201, 213, 4]. Stetter et al. [314] find that conditioning on the mean field (as a summary of the remaining variables) works reasonably well for calcium imaging data at intermediate activity levels, perhaps because such intermediate activity levels are large enough to see coherent source–target effects but not large enough to corrupt the data with too many redundancies. Alternatively, one can perform a standard pairwise approach and then pruning based on survival of conditional testing on every other variable (Wu et al. [365]) or groups of pre-selected parents (Runge et al. [291]), yet these approaches may over-eliminate connections due to redundancy in the network and, crucially, contain no mechanism to capture synergistic contributions.

In contrast to these approaches focussed on redundancy only:

Key Idea 38: Iterative *or* greedy *approaches with conditional transfer entropy can both capture synergies* and *eliminate (only non-required) redundancies [200, 85, 315, 213].*

These iterative approaches gradually build up the set of parents for a given target, whilst only conditioning the TE for new candidate sources on the set of previously selected parents. There are subtle differences between the techniques presented in [200, 85, 315, 213]. Lizier and Rubinov [200] and Stramaglia et al. [315] employ statistical significance calculations to determine stopping conditions (as described above[5]), in contrast to selecting a fixed-size set of parents. The use of statistical significance testing becomes even more important for such multivariate measures than for simple pairwise TE, since it provides an automatic brake on the increasing dimensionality (as more sources are included in the set of parents) as this gradually exhausts the limited statistical power of a given finite data set. Additionally, Faes at el. [85] add the use of non-uniform embeddings for the variables in the TE calculation; Stramaglia et al. [315] use interaction-information-based measures rather than conditional transfer entropies directly, which treat redundancies and synergies in a different manner from that described above; and Lizier and Rubinov [200] also add other optimisation steps including pruning.

[5] Although shuffling the target is used to create surrogates in [315], which is not recommended as per footnote 4.

Open Research Question 8: *Which of the above techniques, a mix of them, or additions to them will prove most convincing for inferring effective connections, whilst eliminating redundancies, capturing synergies, and adapting to the size of available data sets?*

Finally, we note the subtle point that:

Key Idea 39: Iterative *or* greedy *approaches with conditional transfer entropy infer an effective network in which a directed link indicates that the source is* a *parent of the target, in conjunction with the other parent nodes.*

It does not necessarily imply that a parent source provides any unique directed pairwise information to the target. The exact nature of that parental relationship, be it a unique or redundant direct pairwise influence or otherwise mediated with other parents, can be interpreted by examining various pairwise and conditional or higher-order transfer entropies.

7.2.3 Applications of Effective Network Inference

As previously noted, these various TE-based techniques for effective network inference have been widely used in the computational neuroscience domain, as described in [351], having been applied for example to MEG [349, 337], fMRI [187, 211], EEG [315, 213, 85] and spiking neuronal data [142, 323].

Such analysis is not restricted to this domain however, having also been applied to financial market time series [296], supply-chain networks [287], interactions between agents in robotic soccer [62], gene regulatory networks (see Sect. 7.4) and networks in social media dynamics (see Sect. 7.6). We defer detailed description of examples to these subsequent sections.

The crucial output of such effective network inference is an understanding of the directed relationships between the variables, as captured in the sample networks visualised in Fig. 7.2. Further analysis of the resulting network is quite common, for example in looking at how the effective network *changes* as a function of experimental condition, e.g. how cortical effective networks change: between different (auditory) working memory tasks [349], or with increased difficulty of a visuomotor tracking task [187], or in differing levels of consciousness [211]. Algorithms also exist to decipher effective connections not only between variables, but between groups or regions of them [187], e.g. between regions of voxels in fMRI data sets.

Finally, we note that the TRENTOOL software [180, 363] is built from the ground up to provide standard effective network inference via TE for *neural* data,

including some of the graph algorithms described above to address redundancies. Enhancements using conditional TE (as available in JIDT [183]) are planned.

7.3 Applications in Neuroscience

Neuroscience is a fertile ground for innovation in information theory and complex systems. New algorithms for calculating entropy and mutual information come from the efforts to estimate these quantities from sparse data sets. The ever elusive notion of what constitutes complexity has got a little closer to definite capture with Bialek et al.'s idea of predictive information [37].

Not surprisingly the neuroscientists have led the way in the applications of transfer entropy. Indeed, computational neuroscience has arguably been the most important application domain for the transfer entropy. For an in-depth summary of recent research in this area, in particular including exploring effective network inference (see Sect. 7.2) in neuroscience, the reader is referred to the book *Directed information measures in neuroscience* [351]. In this section, rather than canvassing the whole field, we provide a selective summary of certain relevant areas.

Neurons in the mammalian brain communicate by sending voltage spikes to each other, which we can approximate as an irregular series of pulses of fixed height and width. But it turns out that there are serious theoretical issues with calculating the transfer entropy of pulse sequences. Section 7.3.1 takes a look at these deep issues.

It may be advantageous to smooth a pulse sequence in some way or average a number of such sequences and resample. Again, these theoretical issues are only partially resolved. Thus Sect. 7.3.1 delves into the theory without attention to specific applications.

The remaining sections look at neuroscience, which we split into three categories of decreasing temporal or spatial resolution: pulse trains, voxels comparison and EEG (electroencephalography):

Pulse trains (Sect. 7.3.2) may show transfer entropy. We would often like to know if neuron x causes the action of neuron y. Although simple to express, this is a difficult mathematical challenge.

Voxels (Sect. 7.3.3) are the 3D equivalent of pixels. They are obtained with imaging techniques, such as fMRI, PET, MEG.

EEG (Sect. 7.3.3.1) is the lowest spatial resolution, although intermediate in temporal resolution.

7.3.1 TE for Pulse Sequences

Neurons communicate via a continuously variable output voltage or by firing a series of voltage spikes. The first case is easily handled with the methods already

discussed in this book, but most neurons in the vertebrate brain communicate via spike sequences.

A neuron's spike train is a semi-random process, akin to a Poisson point process, with one proviso—there is a refractory period between any two spikes. This presents a real challenge for the TE measures described in Chap. 4. If the pulses do not occur at regular intervals, the concept of previous equally spaced time intervals does not exist. But there clearly is mutual information and transfer entropy between such sequences. Otherwise our brains would not work!

Open Research Question 9: *How can transfer entropy be formulated for irregular pulse sequences or spike trains?*

The most straightforward approach is to interpolate between the spikes, a method which we consider in Sect. 7.3.1.1. The much bigger challenge is to deal with the spikes directly, discussed, albeit briefly, in Sect. 7.3.2.

7.3.1.1 Filtering the Pulses

We can take our series of pulses and extract a range of parameters from it. For several decades, the neuroscience world used the average firing rate of neurons as a measure of their activity. The firing rate as a function of time can thus be considered a continuous variable, and we can use the existing machinery to calculate the information-theoretic quantities.

Unfortunately, it is now recognised that firing rate does not capture all the information in the spike train. At the individual level we know that single spikes are significant. For example, Simon Thorpe and colleagues showed that human rapid discrimination of pictures containing or not containing an animal would have to be based on a single spike travelling from layer to layer through the visual cortex to the infero-temporal cortex where object identification takes place [324, 84].

Bialek et al. [38] found that interpolation by a linear filter could work well. They studied the H1 (motion) neuron in the fly. With a continuous input stimulus, it is possible to measure how much error occurs in the spike representation. If the spike output of the neuron is processed by the correct linear filter, a very high proportion of the information in the input signal can be extracted.

Thus for the H1 neuron at least, we can convert the irregular sequence of spikes to a continuous signal with very little loss of information. We can then study information flow based on continuous signals. However, the fly neurons may not be representative of all neurons, particularly the smaller unmyelinated neurons of the cerebral cortex.

7.3.2 Direct TE Estimation Between Spiking Neurons

Spike trains produced by spiking neurons are naturally represented as (multivariate, jointly distributed) *point processes* [71], continuous-time stochastic processes (Sect. 4.7) where discrete events occur at randomly distributed time intervals [46, 49].

Estimation—and, indeed, a definition—of TE for point processes requires some special treatment. A point process may be specified by a *counting process* $N(t)$, where $N(t)$ denotes the number of events that have occurred in the time interval $(0,t]$. A simple approximation is simply to construct a *discrete-time* process by counting events in time segments of fixed length Δt. Transfer entropies may then be estimated as for discrete processes in general (e.g. Ito et al. [142] estimate TE from spiking neural processes by constructing binary time series for whether spikes were observed or not for each process within 1 ms bins). Improvements may sometimes be obtained by convolving the discretised process with a smoothing kernel, often taken to be Gaussian or half-Gaussian [295, 237, 305]. However, if Δt is too small, so that the number of events in segments is frequently low (or zero), then the resulting discrete time series will be irregular and transfer entropy estimation will likely not be amenable to the various discrete-time techniques; to exacerbate this problem, long history lengths will be required. If, on the other hand, Δt is larger than the typical time scale of lagged feedback within the processes, it will fail to reflect information transfer accurately. Thus the technique may work more or less well in practice, depending on the nature of the processes.

A promising and more principled approach has appeared in recent years (mostly by Brown and colleagues [48, 49, 248, 328, 238, 157]) which deploys parametric Granger causality-like methods in a ML framework (*cf.* Theorem 4.2) to infer directional coupling in spiking neural systems. The starting point for the theory is that a (jointly continuous multivariate) point process is completely specified by its *conditional intensity function* (CIF), a history-dependent generalisation of the rate function for a Poisson process, defined for the ith process as

$$\lambda_i(t|h(t)) \equiv \lim_{dt\to 0} \frac{1}{dt}\mathbf{P}(N_i(t+dt) = N_i(t)+1 \,|\, h(t))\,, \tag{7.1}$$

where $N_i(t)$ is the ith counting process and $h(t) = \{h_i(t)\}$, where $h_i(t) = \{0 < t_{i,1} < t_{i,2} < \ldots < t_{i,J_i} \le t\}$, denotes the history of the ith process (comprising J_i events) up to time t; that is, $t_{i,k}$ is the timestamp of the kth event for the ith process. The ith CIF thus denotes the limiting probability that an event occurs for the ith process in the next time increment, given a prior history for all processes.

It may then be shown [71, 46], that the log-probability of a realisation on $(0,T]$ for a (univariate) counting process $N(t)$ is given by

$$\ln p = \int_0^T \ln \lambda(t|h(t))\, dN(t) - \int_0^T \lambda(t|h(t))\, dt. \tag{7.2}$$

This may be extended for multivariate counting processes $N_i(t)$ [143, 331]. Given a parametrised model $\lambda_i(t|h(t);\boldsymbol{\theta})$ for the CIFs, we thus obtain the log-likelihood function for the model, which is frequently taken to be a generalised linear model (GLM) [46, 328, 248]. In [157] this framework is engaged, with a null hypothesis which excludes source process j from the historical dependencies of target process i, to define a log-likelihood ratio measure described by the author explicitly as *Granger causality*. Regarding non-parametric transfer entropy, we may simply apply Definition 4.10 to the counting processes $N_i(t)$. Then, since the continuous-time transfer entropy defined there is just a (scaled) limit of discrete-time transfer entropies, Theorem 4.2 should (under some regularity conditions) apply in the limit, and again we obtain an asymptotic equivalence for point processes, between parametric ML-based Granger causality and non-parametric transfer entropy.[6].

7.3.3 TE in Brain Imaging

A particularly strong application of TE in neuroscience has been in analysing information flows in brain imaging data sets. As previously pointed out, the reader is referred to the book *Directed information measures in neuroscience* [351] for a detailed overview of such applications. Here, we briefly discuss one application of TE to analysing EEG recordings.

7.3.3.1 TE in EEG

EEG is an old brain imaging method, but it is continuing to grow in importance. It consists of attaching electrodes to the scalp and recording electrical activity therefrom. It might seem quite remarkable that 100 million neurons should produces any sort of interesting or usable behaviour on the scalp. What makes EEG of continuing interest, now stretching for example into the computer games arena, is the advancement of both technological and analysis techniques. From the technology perspective, electrodes have been getting better, picking up stronger signals with less and less effort and pre-preparation of the scalp. Improved electronics has facilitated increased size of the arrays and therefore spatial resolution.

EEG has always had good temporal resolution, but its spatial resolution is dreadful. However, powerful signal processing techniques are improving it all the time. Transfer entropy is the latest technique to come into play. One can distinguish three application areas:

1. Information flow from sensory modalities and external stimuli into the EEG signal (Sect. 7.3.3.2);

[6] In fact we suspect that the equivalence may be *exact* in this case, rather than asymptotic, but we do not as yet have a rigorous proof.

2. Information flow between the components (or electrodes) of the EEG, or beamformed sources within the brain.
3. Phase transitions in the signal. This is a quite different application, relevant, for example, to the onset of an epileptic fit.

We briefly discuss an example of the first application area in the following.

7.3.3.2 Information Flow to EEG from External Stimuli

A relevant example here is the work of Madulara et al. [210] in studying information flow from vision to EEG signals. They test two conditions, eyes open and eyes closed. Not unsurprisingly a large transfer entropy is found between electrodes associated with the occipital lobe (where visual information enters the brain from the eye) and the frontal areas of the brain. The information flow is two way: the net information flow (subtracting one direction from the other) occurs from the frontal areas out to other areas. The value of this net information measure is debatable. Since the brain is highly recurrent, the information flow in both directions is significant. Although an important methodological paper, the results are not surprising. One omission, though, is the conditioning out of other variables (as pointed out in the neuroscience domain by [335, 195, 332, 316]). What effect this would have in this particular experiment is an open question, and indeed as pointed out in Sect. 7.2, conditioning out all other electrodes is generally numerically implausible for realistically sized data sets.

Open Research Question 10: *What happens to EEG transfer entropy after conditioning out other electrodes for each electrode pair?*

7.4 Information Transfer in Biochemical Networks

Fernández and Solé [89] observe that "biological entities perform *computations*", and the key difference between biological and physical systems is the evolutionary payoff associated with information processing in biological systems, e.g. a better ability to "cope with environmental uncertainty". These authors refer in particular to computation taking place in *cellular networks* within an organism, comprising: "the *genome*, in which genes can affect each other's level of expression", "the *proteome*, defined by the set of proteins and their interactions by physical contact", and "the metabolic network (or the *metabolome*), integrated by all metabolites and the pathways that link each other". While these networks are intertwined, they can be considered and modelled separately—for example, one can view genes as *nodes* in gene regulatory networks (GRNs), with expression levels of the genes associated

with each node, directed edges representing the manner in which expression levels of one gene affect another, and other inputs being from environmental factors or other cellular networks. Indeed, the use of Boolean networks[7] to model GRNs is a classic example here [154], where expression levels are binned or discretised as simply *on* or *off*, updating as Boolean functions of the parent genes, and attractors of the network are interpreted to represent cell types.

The aforementioned interpretation of the networks as *computing* a response to their environment has driven interest in information transfer in cellular networks, with such transfer seen as key to understanding the gene *interactions* underpinning the emergent time-series behaviour of the network as a whole [89, 329, 252, 73, 220]. At an abstract level, this is embodied in suggestions that genetic networks occur more naturally near the *edge of chaos* [174] with a balance of ordered and chaotic dynamics optimising computational capabilities near second-order phase transitions [174, 154].[8] More concretely, a key role of information transfer or signalling in biological networks is surmised by Fernández and Solé [89] as nonlinearly generating coherent structures, similar to those we previously described in cellular automata (Sect. 5.1) and swarms (Sect. 5.5), i.e.: "local events involving a few molecules can produce a propagating cascade of signals through the whole system to yield a global response". Clearly:

Key Idea 40: *There is significant potential for transfer entropy to produce key insights regarding the time-series dynamics on biochemical networks—measuring predictive effects of one gene on another, modulation of such effects over time, and indeed inferring effective networks.*

While it is clear that information theory "is the theoretical framework that provides the necessary mathematical tools for the analysis of biological information processing" [220], there are significant issues preventing wide-ranging application of transfer entropy in this domain. Time series of genetic activity levels can be obtained using microarrays (e.g. see [329]), yet it is difficult to obtain time series of adequate length for TE estimation. These difficulties are compounded if one attempts to adequately sample the past state of a target (using standard embedding of k previous samples) or condition on other variables (as discussed by Tung et al. [329] regarding multivariate extensions of TE; i.e. see Sect. 4.2.3). Furthermore, the utility of many microarray data sets for transfer entropy analysis is further compromised in that the time-series observations are *irregularly* sampled (meaning the $\{\boldsymbol{y}_n^{(l)}, x_{n+1}, \boldsymbol{x}_n^{(k)}\}$ tuples are not properly comparable), or taken from periodic attractor states (where TE, if past states are properly embedded, would be trivially zero).

[7] Or indeed random Boolean networks (RBNs), as a class—see the application of transfer entropy to RBNs in Sect. 5.3.

[8] See Sect. 5.3 and Sect. 5.2 for quantitative studies of TE during such phase transitions.

Open Research Question 11: *How can transfer entropy be computed for irregularly sampled time series? For example, using kernel methods and re-sampling techniques to pre-process the data [38].*

Despite these difficulties, Tung et al. [329] have performed a TE analysis on a human cell cycle microarray time-series data set, consisting of 74 samples. This is certainly at the lower end of the amount of data one would require for transfer entropy analysis. TE calculations are made on maximum-entropy binning of the data, with embedding dimension or history length $k = 1$. Tung et al. describe their application as drawing a *causal network*, though it should properly be understood as effective connectivity analysis. Their approach is almost the same as standard pairwise effective connectivity approaches for brain imaging data as described in Sect. 7.2, however Tung et al. assess statistical significance of TE measurements against surrogates obtained by perturbations of the target rather than source time series.[9] They also added a heuristic method to differentiate between direct and indirect connections, by pruning weakest links in potential common driver scenarios (see other approaches discussed in Sect. 7.2). Tung et al. report encouraging results from the experiment, with support found in the scientific literature for around half of the identified directed links.[10]

The use of *models* of biological networks is an approach with the potential to generate a large enough amount of data for practical TE studies. Pahle et al. [252] take a mixed approach, by stochastically coupling experimentally-measured calcium signals to simulated target proteins. Calcium signalling pathways are of interest since a variety of different stimuli (e.g. hormones or nucleotides such as adenosine triphosphate [ATP]) trigger calcium responses while a variety of targets (e.g. proteins and transcription factors) depend in turn on calcium signals. This implies that "specific information is likely to be encoded in the calcium signal", with proposals that it may be encoded in amplitude, frequency, duration, waveforms or timing of oscillations, and that such information is likely to be "decoded again later on" [252]. Pahle et al. generated bivariate time series of calcium and a calcium-dependent enzyme using their mixed-mode modelling, and analysed these with transfer entropy in order to generate new insights into calcium signalling under different conditions. The calculations were made on time series of 10 000 samples, using (box) kernel density estimation, with embedded history length $k = l = 1$. Pahle et al. found that minimum numbers of calcium particles were required to generate significant flows of information, and then the information transfer increased with such numbers (over some range). They also reported that information transfer increased with the (qualitative) complexity of the dynamic mode of the calcium signal, suggesting further

[9] Perturbing the target time series instead of the source is not advisable, since it destroys the relationship between the past state and next value of the target time series (which a null hypothesis test should preserve).

[10] The unvalidated remainder may be as yet unknown interactions, perhaps too weak or unimportant to be identified by previous studies, or could indeed be false identifications.

investigation regarding whether this information is carried in specific properties of the signal (i.e. amplitude, frequency and duration as suggested above). There are interesting parallels here to the work of Laughlin et al. [177] in determining the cost per bit of information, in terms of numbers of ATP molecules in the blowfly visual system. This prompts the question:

Open Research Question 12: *Can we determine direct relationships between transfer entropies in biochemical networks and metabolic costs in the system?*

Similarly, Damiani and Lecca [73] use transfer entropy to infer the effective network underpinning the metabolic reaction network of gemcitabine (an oncological drug), using model-generated data. They use embedded history lengths $k = l = 1$ for the pairwise TE computations, and while considering the use of statistical significance calculations to infer the edges in the network, they in fact use comparison of the TE values against a fixed threshold. Damiani and Lecca evaluate the results against expected and unexpected interactions. They report that thresholded, pairwise TE used alone was overly sensitive—we should not be too surprised by this result given the use of threshold rather than statistical significance tests here, and no use of conditional variables to remove redundancies between potential sources here (see Sect. 7.2). Regardless, Damiani and Lecca suggest that apparent TE could be useful as a "hypothesis generator, whose outcomes may suggest new experiments". Importantly, they subsequently add a step to prune edges from the network using a (model-based) parameter-estimation method on the resulting network to detect non-plausible links. Damiani and Lecca conclude that their model-based pruning step resulted in large improvements to sensitivity and accuracy.

Furthermore, we note the transfer entropy based investigation by Banerji et al. [17] on data from the human primary naive CD4+ T cell intracellular signalling network of 11 proteins, as generated by Sachs et al. [293]. The measure introduced and used by Banerji et al., *network transfer entropy*, is inspired by TE but is model based (considering a stochastic flow type of dynamics). Furthermore it relies on *interventions* or perturbations to infer causal effect, in a similar fashion to Ay and Polani's causal information measure [11] (as discussed in Sect. 4.2.2.1), but without the foundation of Pearl's formalisms [261]. This renders its meaning subtly different from our interpretation of information transfer. With that said, their TE-related analysis of this data set (where perturbations were produced using reagents to activate or inhibit particular proteins) provided novel insights into the network. These included that inhibition of a source protein PIP2 leads to increased flow from other sources to proteins that PIP2 normally activates, which was suggested to indicate the existence of compensatory mechanisms in the network.

Finally, a view to the future here is provided by recent work from Walker et al. [343, 158] who pose the question:

Open Research Question 13: *What informational features "distinguish biological networks from other classes of complex physical systems"?*

From a study of Boolean models of two biological regulatory networks, the authors report that the biological networks exhibit significantly higher levels of transfer entropy as compared to (surrogate) random networks, and networks with in- and out-degrees retained for each node but the network structure randomised (referred to as "scale-free" networks). They also report that the biological networks are distinguished by a "control kernel" exhibiting higher information storage, as well as high information transfer both to and from other nodes. These interesting first results will pave the way for further work to address this crucial question.

In summary, as our data collection and model building capabilities for biochemical networks improve, these applications provide a solid proof of concept that transfer entropy will be a key analysis tool in this domain. With that said, the examples again demonstrate the importance of careful application of TE and its related algorithms.

7.5 Information Transfer in Embodied Cognitive Systems

We use the term *embodied cognitive systems* to refer to Artificial Life, modular robotic, swarm, sensorimotor or multi-agent systems, whose intelligent behaviour emerges out of the interaction between the *brain* or controller, *body* and *environment* (i.e. the *embodiment* of that brain), where the environment often includes similar embodied agents. It is apparent that transfer entropy could play a key role in characterising such systems, in particular in examining information flows between actuators and sensors through the environment, or between distributed agents in the system (whose interaction network provides a key component of the system's embodiment).

Impetus for the study of the key role of information processing *in general* in this area began with early studies of information-theoretic trends with respect to evolutionary time in such systems, e.g. increases in complexity in Artificial Life systems [1, 367], and increases in excess entropy [69] in modular robots [272]. Later, studies turned to examine the role of information in driving the evolution or adaptation of such systems—an approach referred to as *guided self-organisation* [268, 269]. For example, Sporns and Lungarella [312] evolved hand–eye co-ordination to grab a moving object using maximisation of a measure of neural complexity,[11] and demon-

[11] The measure of neural complexity used is known as Tononi–Sporns–Edelman (TSE) complexity [326]. It is an information-theoretic measure which seeks to measure complexity as a balance between integration and segregation of components in a multivariate time series. The measure builds on sums and differences of multi-information or integration terms (see Sect. 3.2.2.2, taken over collections of neural variables.

strated that this solution contained more intrinsic diversity than solutions from task-driven evolution; the increased diversity may afford greater flexibility to the system. Prokopenko et al. [271] evolved fast-moving snakebots using maximisation of an information-theoretic measure of co-ordination, while Martius et al. use the same measure to drive adaptation of various modular robotic systems and observe high behavioural variety [217]. Also, Sperati et al. [310] observed interesting periodic behaviour and complex structure in groups of robots which were evolved to maximise their mutual information.

More specifically, Prokopenko et al. suggested a key role of information transfer in modular systems in underpinning co-ordinated motion (i.e. communications between modules keeping them co-ordinated) [271]. Further work implying the importance of information transfer in embodied cognition includes the concept of *empowerment* [161, 162], which suggests that maximising self-perception of influence over the environment (measured as channel capacity between an agent's actuators and sensors through the environment) is a useful intrinsic selection pressure. Several related studies have sought to characterise causal information flows in the sensorimotor loop of embodied agents [264, 10].

In this vein, we summarise several studies applying transfer entropy to study information flows in such systems. A seminal application of TE in this area was by Lungarella and Sporns [206]. They hypothesised that:

Key Idea 41: *Sensorimotor interaction and morphological structure induce information structure in the sensory input and neural system, promoting information processing and flow between sensory input and motor output [206]* which can be quantified by transfer entropy.

Lungarella and Sporns produced a comprehensive study, measuring transfer entropy in both directions between visual sensors and movement actuators in a variety of embodied systems (i.e. a visual system tracking a moving ball, a four-legged robot moving amongst blocks, and a wheeled robot moving amongst spheres). Transfer entropy was measured from experiments with on the order of 1000s of (discrete) time steps, by binning the data in eight bins, with history embedding length $k = 1$. Lungarella and Sporns demonstrated that in their systems:

1. Information flow is spatially and temporally specific, with different parts of the visual field driving and experiencing different effects over different time lags. This is revealed more explicitly by investigations with the local transfer entropy discussed below [192], and aligns with many other studies, including studies of the distribution of TE in neural networks [215, 321]. In particular, information flow was found to increase with the amount of environmental changes causing behavioural responses, as would be expected.
2. Information flow can be affected by learning. By inducing rewards and aversion in their systems, they found for example that larger information flows resulted from visual sensors responding to the colour being rewarded.

3. Information flow can be affected by changes in body morphology. They demonstrated larger information flows when receptive fields of visual sensors were larger for more central receptors in the retina.

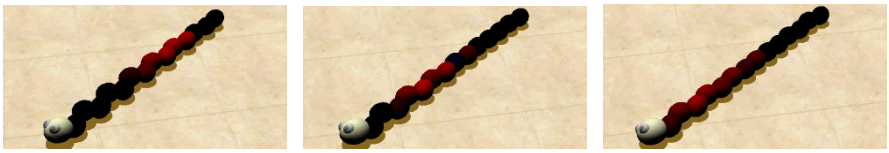
Lungarella and Sporns interpret their results as suggesting a fundamental link between physical embeddedness (embodiment) and information structure.

Lizier et al. [192] introduced two key innovations for the application of transfer entropy in this domain. The first was to invert the usual use of transfer entropy, applying it for the first time as a fitness function in the evolution of adaptive behaviour, as an example of guided self-organisation. The second was to switch focus to using the *local* transfer entropy [195] (as introduced in Sect. 4.2.5) to characterise the dynamics in the system on a local scale in space and time. Lizier et al. focussed on a snakebot—a snake-like robot with separately controlled modules along its body, whose individual actuation was evolved via genetic programming (GP), whilst the actual motion emerged from the interaction between the modules and their environment. The approach altered the GP to maximise transfer entropy between adjacent modules, measuring TE using kernel estimation. While this approach did not result in a particularly fast-moving snake (as had been hypothesised), it did result in coherent travelling information waves along the snake, which were revealed only by local transfer entropy (as shown in Fig. 7.3). These waves are akin to gliders in CAs [195] and cascades in swarms [345]—both in terms of raw dynamics and coherent TE characteristics. Other natural systems (where TE has not yet been studied) appear to exhibit very similar coherent propagating structures (e.g. waves of opening and closing of stomatal apertures in plants [260], and waves of huddling motion in Emperor penguins [371] etc., as discussed in Sect. 5.5). The results suggest that:

Key Idea 42: *Such coherent wave structures may emerge as a resonant mode in evolution for information flow.*

While it is possible that these wave structures may not transfer the most information between individual units (in all such systems), they appear robust and optimal for coherent communication over long distances, and may be simple to construct via evolutionary steps. Revisiting Lungarella and Sporns' finding above [206] that TE can be affected by learning, the snakebot experiment explicitly shows that TE can *shape* learning.

In a similar vein, Nakajima et al. [235] explored the use of local TE [195] to characterise dynamics in robotic platforms. They investigated a soft robotic arm platform (both a simulator and physical platform) using local TE and local complete TE (see Sect. 4.2.3). The local TE values are estimated using symbolic transfer entropy [313] (see Sect. 4.3.2), with a history embedding length of $k = 3$ past values used in the symbols. Nakajima et al. demonstrated that the local TE clearly reveals the effect of external damage to the robotic arm, visualising waves of impact travelling along the arm, followed by waves of corrective motion and their collision. These local information waves are similar to those observed in [192] above. Naka-



(a) (b) (c)

Fig. 7.3 Snakebot modules coloured to indicate incoming local transfer entropy (black is 0.0 bits, red is 2.8 bits) from the neighbouring module toward the tail, for three consecutive time steps. The information transfer wave from the tail appears to communicate a straightening behaviour here. Videos of these coherent cascades of local transfer entropy in the evolved snakebot are available at http://youtu.be/HmRI5hfaBQ8 (NB: This figure was first published in [192])

jima et al. also demonstrated a perhaps surprisingly large robustness of the local TE measure to noise in the simulator dynamics. This robustness would be assisted by the long time series available for analysis here (30,000 steps for the simulator), though that effect was not quantified. The robustness is an important result because of the noise inherent in physical robotic applications, and indeed the application to the physical robot was upheld by results similar to the simulator.

Williams and Beer [357] also sought to characterise temporal behaviour or dynamics of embodied agents using transfer entropy. Their experiments involved an agent moving along a horizontal line while observing circular objects which fall toward it from above. The agent observes the objects with seven differently angled distance-measuring rays, which are inputs to a five-node (three interneuron, two motor neuron) recurrent neural network controlling the agent's movement. The parameters of the agent were evolved to observe two falling objects, and then to minimise its distance to the second if it is smaller than the first, or to maximise that distance if the second object is larger. Partially localised mutual information (i.e. specific information, see Sect. 3.2.2) and transfer entropy, with $k = l = 1$, was measured between the specific size of the second object and the individual activation levels of the sensors and three interneurons, as a function of time.[12] The observations were discretised into 100 bins. The approach revealed the gradual appearance and decay of information about the object size in each sensor (explained by the sensor's orientation). Furthermore, the measures revealed when, and on which neurons, information about the object size appeared, and how this differed as a function of the object size. Crucially,

Key Result 12: *This approach using transfer entropy revealed how information was distributed* spatially *and* temporally *in the system, allowing a precise description of how the embodied computation took place in the agent.*

[12] Note that this is subtly different from the *local* transfer entropy as described earlier, using an ensemble approach to measure average transfer entropy on observations taken only at a given time step over many repeat trials (similar to the ensemble approach described in Sect. 4.3.1.1.)

These applications, and others still (e.g. application to a chaos-controlled robot by Lungarella et al. [205], and to analyse evolution of neural networks in an Artificial Life simulator by Lizier et al. [189]), demonstrate the ability of TE to provide deep insights into embodied cognitive systems, and suggest strong potential for it as a tool in future development of emergent intelligent systems.

These investigations suggest several open research questions. While transfer entropy has clearly provided useful insights in these systems, its application has been rather specifically tailored to each situation:

Open Research Question 14: *What are the more important information channels to focus on regarding information flow in embodied cognitive systems—between nodes in an agent's neural network, from actuators to sensors through the environment, or between distributed agents in the system?*

We wonder whether any particular application types or recipes for applying transfer entropy in this domain will emerge as a generally useful approach to delivering insights here, for example:

Open Research Question 15: *Are there characteristics in the dynamics of transfer entropy that can be linked to key evolutionary or adaptive steps in an embodied agent's development?*

And finally, there is the engineering question of whether transfer entropy can be used for information-driven design of application-specific useful behaviours:[13]

Open Research Question 16: *Can transfer entropy or other measures of information dynamics be utilised as an application-independent, intrinsic goal to drive the guided self-organisation of embodied cognitive systems, via adaptation or evolution? For which types of behaviour would this provide a useful template (e.g. top-down causation [342])? How could the intrinsic capability conferred by guiding for high transfer entropy then be built on to produce application-specific utility?*

[13] See also e.g. the use of transfer entropy to guide the adaptation of echo state neural networks [245], and related work using active information storage to guide adaptation of recurrent neural networks [74].

7.6 Information Transfer in Social Media

Social media, including Facebook and Twitter, are transforming the manner in which people interact in their daily lives. And interestingly for researchers, their digital footprint allows unprecedented documentation of the structure and dynamics of such interactions. Many studies are taking advantage of this opportunity (e.g. [90, 76]), and this includes the use of transfer entropy of course. Its inherent ability to quantify the predictive effect of one variable—i.e. the actions/comments of a person here—on another means that it is an obvious candidate to explore, for example: how different users respond to stimuli from others, which people have the strongest predictive effects on others in such social networks, how this relates to social network structure, and how such interactions relate to real-world ties.

Ver Steeg and Galstyan [334] analyse transfer entropy between the contents of status events (*tweets*) of various users' accounts on the Twitter social network. They analysed a data set of one month of tweets from 770 interacting users (with at least 100 tweets each), computing TE between each pair. The first step in their analysis was to pre-process the tweets into a lower-dimensional space than the original 140-character strings. The processing involves learning a set of topics (*topic model*), then each tweet is turned into a vector of scores on each topic (a *topic vector*). It is these multi-dimensional values that TE was computed on. Next, the usual TE approach was altered to consider the point-process-like nature of tweets. For a given source–target account pair, Ver Steeg and Galstyan constructed $\{$`source`, `target`, `target_past`$\}$ tuples for estimating the PDFs by taking each tweet by the target and combining these with the most recent *previous* tweets of the source and by the target. TE was then computed on the processed data by way of a KSG estimator (with three nearest neighbours, see Sect. 4.3.1). For faster computational speed, calculation of TE was made over subsets of 100 tweet exchanges, then averaged over available subsets of this size. Ver Steeg and Galstyan thresholded the TE values in order to infer an effective network structure (Sect. 7.2), indicating social interactions amongst users. They then compared these inferred links to the underlying social network, demonstrating two perhaps surprising findings: (i) that many of the most predictive links were simply not present in the social network (perhaps explained by unseen common drivers), while (ii) most of the social network links did not have high transfer entropy. Ver Steeg and Galstyan then used a local transfer entropy analysis ([195], see Sect. 4.2.5; using KSG estimators here) to explore which tweet exchanges *contribute most* to TE for various pairs, using this to reveal some unseen common drivers; for example various news site feeds were identified as common drivers of pairs of users. These common drivers could perhaps be handled using conditional TE approaches (see Sect. 4.2.3 and Sect. 7.2) if their time series were included in the analysis. Finally, Ver Steeg and Galstyan found that strong transfer

entropy is a statistically significant predictor of *mentions* on Twitter[14] adding further interpretable meaning from the measure.

Oka and Ikegami also used transfer entropy to explore dynamics in social media and web searches [246, 247], though in contrast to exploring the relationships between users, they explored the relationships between keywords via the time-series dynamics of their Google search and Twitter mention volumes. Oka and Ikegami manually selected 26 keywords in [246] and 300 keywords in [247] covering *bursty*, *bursty with chatter* and *chatter with rare burst dynamics*, gathering time series of daily Google search volumes and Twitter keyword volumes over 3 months [246] and 11 months [247]. They also randomly selected 126 keywords in [246] and 1000 keywords in [247] to analyse TE in keyword volumes within Twitter only (per hour data). Transfer entropy was calculated by computing each of the underlying joint entropies using permutation entropy [16] (which appears to make this calculation equivalent to a symbolic transfer entropy [313], see Sect. 4.3.2). The calculation used $k = 1$ step of history, though suggests others may have been checked. Furthermore, the calculations in [246] used a sliding time window of 18 days, allowing some interpretation of the temporal TE dynamics. Oka and Ikegami found generally a larger transfer from Twitter to Google, perhaps suggesting precedence of information about real events on the Twitter network. They then provided an interesting interpretation of their results, following suggestions of *default* and *reactive* modes in cognition. They reported that:

Key Result 13: *Inner TE activity in Twitter becomes suppressed when transfer from Google is high, then increases as such incoming flow reduces (suggesting activation of default mode activity following reaction to stimulus).*

Additionally, they reported that more frequent keywords typically have larger outgoing transfer to less frequent keywords, suggesting the internal self-sustained driving of such default mode activities.

Considering edits to the free online encyclopaedia Wikipedia, Bauer et al. [32] examined whether the temporal behaviour of editors was predictive of whether such editors were connected in a social network. They applied TE analysis to time series of the number of edits made each day by each editor, with a binary discretisation or binning applied at various thresholds. The analysis used $k = 5$ past history values of the target time series. Bauer et al. selected editors for the analysis simply from those who had edited sample Wikipedia pages ("Elvis Presley"—1963 users, and "Anarchism"—1218 users). They then translated the pairwise TE measurements into an undirected effective network (i.e. applying effective network analysis, see Sect. 7.2), by taking the maximum TE in each direction for a pair of editors, then using a variable threshold to determine presence of links in the network. The effec-

[14] *Mentions* being where one user is explicitly named in the status or *tweet* of another. At first this may seem a trivial result, however TE here was based on related content, not mentions; the result is that mentions are correlated to strong predictive influence of content.

tive networks were then compared with a social network, constructed from whether editors directly interacted in their personal Wikipedia *Talk* pages (used for direct discussion between editors). They point out that this method cannot capture the entire social network between the editors, but provides a *lower bound*. To interpret the results, they varied the threshold for inferring the effective connections in the TE network, and evaluated precision and recall of the social network as a function of this threshold (creating an ROC—receiver operating characteristic—profile). Bauer et al. reported that they obtained precision (proportion of inferences which are correct) at approximately 20 times better than random guessing for a given level of recall (proportion of true links which are inferred), and emphasised that this should be seen as a lower bound on the performance of the algorithm (given that the underlying social network is only partially known). They concluded that:

Key Result 14: *If the time series of edits of a source editor on Wikipedia is predictive of edits by a target editor (as measured by TE), then this is a useful implication of whether the two actually interact [32].*

These early results are indicative of strong potential for the use of TE in data mining here, as the volume of data produced by social networks continues to grow. There is much scope for further investigation here (e.g. see also [45]), in terms of both the community settling on analysis approaches, and further applications. We have seen several types of pre-processing here, and several types of information channels (between users' content, their activity timings, and even between activity volumes on different websites):

Open Research Question 17: *On which information channels in social media networks will transfer entropy prove to be most revealing of underlying structure?*

Open Research Question 18: *Given high dimensionality, and limited samples per user, how should one pre-process social media data in order to best capture the relevant information and yield to transfer entropy analysis?*

7.7 Summary

In this chapter we have canvassed the wide-ranging use of transfer entropy in a suite of fields, focussing on biochemical networks, embodied cognitive systems, social

media, neural and other physiological data, as well as general algorithms for inferring effective network structures from multivariate time-series data. These examples showed the flexibility of the measure in being subtly adjusted to accommodate the needs of each application, while providing novel insights in each domain. It is clear that transfer entropy will have a growing impact in each of these fields, and the manner in which it is flourishing in those suggests its continued growth into other new domains. It is also clear from these studies that exceptional care is needed in data processing and TE estimators to achieve robust and reliable results.

Chapter 8
Concluding Remarks

This book has surveyed an important but still emerging field. We have noted some open research questions along the way, but now will have a brief look into the future. The first issue (Sect. 8.1) is estimation from real-world data sets. Progress in this area is essential for wider application. The second is gaining a deeper understanding of the differences between different forms of transfer entropy, especially the significance of the global measure found to be somewhat important in the Ising model and possibly in other forms of phase transition (Sect. 8.2). But we end the book with a deep theoretical issue—the link between energy and transfer entropy (Sect. 8.3).

8.1 Estimation

Much of this book has been about theory and canonical systems. If we want to empirically calculate the transfer entropy for a spin system, cellular automaton or any of these other abstract systems, all we need is computer power. We can simply gather as much data as necessary to get the required statistical robustness.

The situation is rather different for real-word systems. Data is often limited, and we may have little control over its sampling. In stock markets, briefly mentioned in Chap. 6, the data is defined by as and when sales occurred. Estimation may be either parametric or non-parametric.

8.1.1 Non-parametric Estimation

We saw in the discussion of mutual information that there are numerous estimators, and there is no one perfect estimator. Which one performs the best depends on the data set. Thus some empirical investigation will be needed in any given case. Transfer entropy is usually determined as a linear combination of entropies or mu-

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9_8

tual information terms: there is very little work we know of for estimating transfer entropy directly, with the extension of the KSG estimator to conditional mutual information and therefore TE [93, 110, 337, 350] (and see Sect. 4.3.1) being one very notable exception. The extensive discussion on direct mutual information estimation needs to be replicated for TE. So a simple open question is:

Open Research Question 19: *How do the entropy and mutual information estimators perform on different known statistical distributions, especially in cases where the theoretical distribution is known [124, 144]?*

We can then ask whether there are other direct estimators for TE, potentially extending other algorithms for MI, or some other approach as yet unknown.

Open Research Question 20: *Are there additional good non-parametric estimators for transfer entropy which avoid summation of entropic quantities, following the extension of [93, 110, 337, 350] for KSG-style TE estimation?*

Of particular interest here will be extensions of Bayesian techniques for inferring entropies, e.g. [240], to TE.

8.1.2 Parametric Estimation

Almost universal across diverse fields is the drive to find parametric models which simulate observed data. The GARCH models (see Sect. 4.4.3) of economic time series are a good case in point. If we can fit a parametric model, or, even better, have good evidence that a particular model does underlie some system, then the estimation problem is much easier.

At one extreme, we have the maximum-likelihood estimator from Barnett and Bossomaier [23], which not only provides a TE estimate but also yields a statistical test of its reliability. At the other, we have an increasing number of theoretical calculations of TE for different distributions. The recent work by Hahs and Pethel [124] provides such theoretical values for autoregressive models, which should be of considerable value for economic time series.

Somewhat related to parametric estimation is a new direction in causality taken by Sugihara et al. [319]. They point out that the stochastic methods such as Granger causality do not work for deterministic systems, as indeed acknowledged by Granger himself [112]. They propose new methods, which work for coupled chaotic systems, seemingly random, but actually deterministic.

8.1.3 Non-stationary Systems

Many real-world systems are non-stationary, thus their estimation from time-series data runs into difficulties. There is a trade-off between accuracy of estimation of the underlying statistics (long time series) and the need for the series to be short so that it is stationary over the series.

One approach, where multiple time-series samples or trials are available, is an integrated estimator which fuses both time series and ensemble [110] (see Sect. 4.3.1.1). Another possibility is to explore Nason's new wavelet estimators for stationarity [236].

8.2 Systems with Many Variables

When systems have many variables, non-parametric estimation of transfer entropy from empirical data becomes very challenging. The Ising model discussed in Chap. 5 was found to need global transfer entropy, where the information flow is evaluated from all variables except one, collectively, to that excluded variable. One current estimator for this uses a variant, $\mathbf{I}((X_1, X_2, X_3, ...X_{n-1}) : X_n)$ (Eqn. 8.1), of the multi-information (Eqn. 3.14):

$$\mathbf{I}((X_1, X_2, X_3, ...X_{n-1}) : X_n) = \mathbf{I}(X_1 : X_2 : X_3 : \ldots X_n) - \mathbf{I}(X_1 : X_2 : X_3 : \ldots X_{n-1}). \quad (8.1)$$

But calculating this is hard work. Kraskov et al. [168] get good results for dimension up to 8, but using 50,000 points with 100 repetitions! Contrast this with stock market data where, in the Dow Jones Index we have 30 variables, and we would have only around 260 points per year for day-close prices.

The other issue with multivariable systems is avoiding inferring indirect interactions. So, if A causes B and B causes C, A might appear to cause C (also known as a cascade effect or pathway scenario—see Sect. 4.2.3 and Sect. 7.2.2).[1] Several recent proposals have been made in order to remove such redundancies from effective network inference (as well as to incorporate synergies), using conditional TEs as described in Sect. 7.2.2. This conditioning out process is also very computationally and data demanding.

Open Research Question 21: *How can non-parametric estimators for global TE and pairwise conditioning be improved, in terms of efficiency as well as robustness to small data sets?*

[1] Similar common driver effects leading to indirect inferences are also described in these sections.

8.3 Touching the Void: the Link to Thermodynamics

Finally we return to a deep theoretical idea: the link between information and energy. Feynman, Landauer, Bennett and others [172, 173, 34, 275] have pursued an understanding of the energetic costs of computation. For classical systems, the overarching result is that destroying information costs energy, a minimum of $kT\ln(2)$ joules per bit, where k is Boltzmann's constant and T is absolute temperature. kT is like a quantum of energy in quantum physics, a sort of minimal quantity of energy in any system (but it is not indivisible in the manner of quantum mechanics). This limit holds throughout the universe, and no computer anywhere could do better (in classical terms). But we are a very long way from reaching this thermodynamic limit of computation.

The human brain does rather better. When the IBM computer Watson recently beat top human players in the sophisticated TV game, Jeopardy, one thing was down-played. Watson is a supercomputer with several thousand cores and terabytes of main memory. It is physically huge, and consumes a huge amount of energy. The human brain is about 10,000 times more energy efficient at the time of writing.

The brain carries out computation by sending messages between neurons, via voltage spikes in the cerebral cortex. Thus it fuses computation and communication. Just as kT is the effective unit of energy in thermodynamics, in animal physiology the unit is the energy released when a molecule of adenosine triphosphate (ATP) is converted to adenosine diphosphate (ADP). This is the near universal, indivisible energy quantity of much animal physiology, not just brains and neurons. This energy is about $25kT$ at body temperature. Lauglin et al. [177] found that the energy cost of transmitting a bit across a chemical synapse (i.e. between neurons) was around 10^4 ATP molecules. So, the energy cost is about 10^5 kT.

But if we could build a maximally efficient computer, what would its theoretical limit to information communication be, a limit, such as that for destroying information, valid across the entire universe? The recent work by Prokopenko and Lizier [275, 273], referred to in the opening chapter, defines such a limit. One might guess that transferring information from one system to another would cost kT per bit as well. In principle any measure of information flow should thus generate this value. And, indeed it does, as Prokopenko and Lizier have shown. Thus amongst all statistics for this and that, TE links directly to the energetics of computation.

The big open research question with which we end the book asks for experimental tests in real systems:

Open Research Question 22: *Can we relate the energy of communication, in neurons or other systems, to the transfer entropy required of the communication?*

References

1. C. Adami. What is complexity? *Bioessays*, 24(12):1085–1094, 2002.
2. P.-O. Amblard and O. J. J. Michel. The relation between Granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2013.
3. N. Ancona, D. Marinazzo, and S. Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70:056221+, 2004.
4. L. Angelini, M. de Tommaso, D. Marinazzo, L. Nitti, M. Pellicoro, and S. Stramaglia. Redundant variables and Granger causality. *Physical Review E*, 81(3):037201+, 2010.
5. M. A. Antal, C. Böde, and P. Csermely. Perturbation waves in proteins and protein networks: Applications of percolation and game theories in signaling and drug design. *Current Protein and Peptide Science*, 10(2):161–172, 2009.
6. M. Argollo de Menezes, C. F. Moukarzel, and T. J. P. Penna. First-order transition in small-world networks. *Europhysics Letters*, 50(5):574, 2000.
7. V. I. Arnold. *Catastrophe theory*. Springer, 1992.
8. F. M. Atay, J. Jost, and A. Wende. Delays, connection topology, and synchronization of coupled chaotic maps. *Physical Review Letters*, 92(14):144101+, 2004.
9. F. M. Atay and O. Karabacak. Stability of coupled map networks with delays. *SIAM Journal on Applied Dynamical Systems*, 5(3):508–527, 2006.
10. N. Ay and K. Ghazi-Zahedi. On the causal structure of the sensorimotor loop. Technical Report 13-10-031, Santa Fe Institute, 2013.
11. N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–41, 2008.
12. L. Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 1900.
13. R. T. Baillie. Long memory processes and fractional integration in econometrics. *J. Econometrics*, 73:5–59, 1996.
14. P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters*, 59(4):381–384, 1987.
15. M. Baker and J. Wurgler. *Investor sentiment in the stock market*. National Bureau of Economic Research Cambridge, Mass., USA, 2007.
16. C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88(17), 2002.
17. C. R. S. Banerji, S. Severini, and A. E. Teschendorff. Network transfer entropy and metric space for causality inference. *Physical Review E*, 87(5):052814+, May 2013.
18. A.-L. Barabási. Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413, 2009.
19. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
20. A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A*, 281:69–77, 2000.

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9

21. A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288:50–59, 2003.
22. L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.*, 103(23):238701, 2009.
23. L. Barnett and T. Bossomaier. Transfer entropy as a log-likelihood ratio. *Phys. Rev. Lett.*, 109(13):138105, 2012.
24. L. Barnett, M. Harré, J. Lizier, A. K. Seth, and T. Bossomaier. Information flow in a kinetic Ising model peaks in the disordered phase. *Phys. Rev. Lett.*, 111:177203, 2013.
25. L. Barnett and A. K. Seth. Behaviour of Granger causality under filtering: Theoretical invariance and practical application. *J. Neurosci. Methods*, 201(2):404–419, 2011.
26. L. Barnett and A. K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *J. Neurosci. Methods*, 223:50–68, 2014.
27. L. Barnett and A. K. Seth. Detectability of Granger causality for subsampled continuous-time neurophysiological processes. In review, *J. Neurosci. Methods*, Nov. 2015. Preprint: http://arxiv.org/abs/1606.08644, 2016.
28. A. B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E*, 91:052802, 2015.
29. A. B. Barrett, L. Barnett, and A. K. Seth. Multivariate Granger causality and generalized variance. *Phys. Rev. E*, 81(4):41907, 2010.
30. L. Basnarkov and V. Urumov. Phase transitions in the Kuramoto model. *Physical Review E*, 76(5):057201+, 2007.
31. F. Bauer, F. M. Atay, and J. Jost. Synchronization in discrete-time networks with general pairwise coupling. *Nonlinearity*, 22(10):2333–2351, 2009.
32. T. L. Bauer, R. Colbaugh, K. Glass, and D. Schnizlein. Use of transfer entropy to infer relationships from behavior. In *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*, CSIIRW '13, New York, NY, USA, 2013. ACM.
33. C. H. Bennett. The thermodynamics of computation: A review. *Int. J. Theor. Phys.*, 21(12):905–940, 1982.
34. C. H. Bennett. Notes on Landauer's principle, reversible computation, and Maxwell's demon. *Studies in History and Philosophy of Science Part B*, 34:501–510, 2003.
35. N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. Shared information—New insights and problems in decomposing information in complex systems. In T. Gilbert, M. Kirkilionis, and G. Nicolis, editors, *Proceedings of the European Conference on Complex Systems 2012*, Springer Proceedings in Complexity, pages 251–269. Springer, Switzerland, 2013.
36. N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.
37. W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.
38. W. Bialek, F. Rieke, R. d. R. v. Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1854–1856, 1991.
39. S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashions, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.
40. F. Black and M. Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, pages 637–654, 1973.
41. J. Boedecker, O. Obst, J. T. Lizier, Mayer, and M. Asada. Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, 131(3):205–213, 2012.
42. E. M. Bollt. Synchronization as a process of sharing and transferring information. *Int. J. Bifurcation Chaos*, 22(11):1250261+, 2012.
43. E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York, 1999.
44. J. Bonachela, H. Hinrichsen, and M. Muñoz. Entropy estimates of small data sets. *J. Phys. A: Math. Theor.*, 41:202001–202009, 2008.
45. J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, and A. Vespignani. The dynamic of information-driven coordination phenomena: a transfer entropy analysis, 2015. arXiv:1507.06106.

46. D. R. Brillinger. Nerve cell spike train data analysis: a progression of technique. *J. Am. Stat. Assoc.*, 87(418):260–271, 1992.
47. D. R. Brillinger. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, 18:163–183, 2004.
48. E. N. Brown, R. Barbieri, U. T. Eden, and L. M. Frank. Likelihood methods for neural spike train data analysis. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, chapter 9, pages 253–86. Chapman & Hall/CRC, 2003.
49. E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nat. Neurosci.*, 7(5):456–461, 2004.
50. A. Buehlmann and G. Deco. Optimal information transfer in the cortex through synchronization. *PLoS Computational Biology*, 6(9):e1000934+, 2010.
51. J. Buhl, D. J. T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and S. J. Simpson. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406, 2006.
52. S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-Organization in Biological Systems*. Princeton University Press, 2003.
53. T. Carter. Sync model. NetLogo model; accessed July 28, 2014; http://csustan.csustan.edu/%7Etom/Lecture-Notes/Models/NetLogo/Sync/Sync.html.
54. R. V. Ceguerra, J. T. Lizier, and A. Y. Zomaya. Information storage and transfer in the synchronization process in locally-connected networks. In *2011 IEEE Symposium on Artificial Life (ALIFE)*, pages 54–61. IEEE, 2011.
55. C. J. Cellucci, A. M. Albano, and P. E. Rapp. Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Review E*, 71(6):66208, 2005.
56. M. Chávez, J. Martinerie, and M. Le Van Quyen. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. *Journal of Neuroscience Methods*, 124(2):113–128, 2003.
57. A. H. Chen and T. F. Siems. The effects of terrorism on global capital markets. *European Journal of Political Economy*, 20(2):349–366, 2004.
58. P. E. Cheng, J. W. Liou, M. Liou, and J. A. D. Aston. Data information in contingency tables: A fallacy of hierarchical loglinear models. *Journal of Data Science*, 4(4):387–398, 2006.
59. A. Chepizhko and V. Kulinskii. On the relation between Vicsek and Kuramoto models of spontaneous synchronization. *Physica A: Statistical Mechanics and its Applications*, 389:5347–5352, 2010.
60. D. Chicharro and A. Ledberg. When two become one: The limits of causality analysis of brain dynamics. *PLoS ONE*, 7(3):e32466+, 2012.
61. K. Clarke, S. Hoppen, and L. Gaydos. A self-modifying cellular automata model of historical urbanization in the San Francisco Bay Area. *Environment and Planning B-Planning and Design*, 24(2):247–261, 1997.
62. O. M. Cliff, J. T. Lizier, X. R. Wang, P. Wang, O. Obst, and M. Prokopenko. Towards quantifying interaction networks in a football match. In S. Behnke, M. Veloso, A. Visser, and R. Xiong, editors, *RoboCup 2013: Robot World Cup XVII*, volume 8371 of *Lecture Notes in Computer Science*, pages 1–12. Springer, Berlin/Heidelberg, 2014.
63. J. H. Conway. What is Life? In E. Berlekamp, J. H. Conway, and R. Guy, editors, *Winning ways for your mathematical plays*, volume 2, ch. 25, pages 927–962. Academic Press, New York, 1982.
64. M. Cook. Universality in elementary cellular automata. *Complex Systems*, 15(1):1–40, 2004.
65. J.-M. Courtault, Y. Kabanov, B. Bru, P. Crépel, I. Lebon, and A. Le Marchand. Louis Bachelier on the centenary of Théorie de la spéculation. *Mathematical Finance*, 10(3):339–353, 2000.
66. I. Couzin. Collective minds. *Nature*, 445(7129):715–715, 2007.
67. I. D. Couzin, R. James, D. P. Croft, and J. Krause. Social organization and information transfer in schooling fishes. In C. Brown, K. N. Laland, and J. Krause, editors, *Fish Cognition and Behavior*, Fish and Aquatic Resources, pages 166–185. Blackwell Publishing, 2006.
68. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, July 2006.

69. J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos*, 13(1):25–54, 2003.
70. J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 223–269. Addison-Wesley, 1990.
71. D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York, 2003.
72. S. R. X. Dall, L.-A. Giraldeau, O. Olsson, J. M. McNamara, and D. W. Stephens. Information and its use by animals in evolutionary ecology. *Trends in Ecology and Evolution*, 20(4):187–193, Apr. 2005.
73. C. Damiani and P. Lecca. Model identification using correlation-based inference and transfer entropy estimation. In *Proceedings of the 2011 Fifth UKSim European Symposium on Computer Modeling and Simulation (EMS)*, pages 129–134. IEEE, 2011.
74. S. Dasgupta, F. Wörgötter, and P. Manoonpong. Information dynamics based self-adaptive reservoir for delay temporal memory tasks. *Evolving Systems*, 4(4):235–249, 2013.
75. B. de Giusti. Moofushi kandu fish.jpg. used under Creative Commons Attribution-Share Alike 2.5 Italy (CC-BY-SA-2.5-IT); accessed 25 July 2014; https://upload.wikimedia.org/wikipedia/commons/3/32/Moofushi_Kandu_fish.jpg.
76. S. DeDeo. Collective phenomena and non-finite state computation in a human social system. *PLoS ONE*, 8(10):e75818+, 2013.
77. Z. Deng, J. Wu, and W. Guo. Rényi information flow in the Ising model with single-spin dynamics. *Physical Review E*, 90(6):063308+, 2014.
78. M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10:325–340, 1999.
79. W. J. Dobson. The day nothing much changed. *Foreign Policy*, (156):22–25, 2006.
80. J. Doob. *Stochastic Processes*. John Wiley, New York, 1953.
81. S. N. Dorogovstev and J. F. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, New York, 2005.
82. A. W. F. Edwards. *Likelihood (Expanded Edition)*. Johns Hopkins University Press, Baltimore, 1992.
83. A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905.
84. M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scences. *J. Cog. Neurosci.*, 13(2):171–180, 2001.
85. L. Faes, G. Nollo, and A. Porta. Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83:051112+, 2011.
86. L. Faes, G. Nollo, and A. Porta. Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Computers in Biology and Medicine*, 42(3):290–297, 2012.
87. L. Faes, A. Porta, G. Rossato, A. Adami, D. Tonon, A. Corica, and G. Nollo. Investigating the mechanisms of cardiovascular and cerebrovascular regulation in orthostatic syncope through an information decomposition strategy. *Autonomic Neuroscience: Basic and Clinical*, 178(1-2):76–82, 2013.
88. R. M. Fano. *Transmission of information: A statistical theory of communications*. MIT Press, Cambridge, MA, USA, 1961.
89. P. Fernández and R. V. Solé. The role of computation in complex regulatory networks. In E. V. Koonin, Y. I. Wolf, and G. P. Karev, editors, *Scale-free Networks and Genome Biology*, pages 206–225. Landes Bioscience, Georgetown, TX, 2006.
90. E. Ferrara. A large-scale community structure analysis in Facebook. *EPJ Data Science*, 1(1):9+, 2012.
91. D. V. Foster and P. Grassberger. Lower bounds on mutual information. *Physical Review E*, 83(1):010101+, 2011.

92. A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A.*, 33(2):1134–1140, 1986.
93. S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters*, 99(20):204101+, 2007.
94. P. Fries. A mechanism for cognitive dynamics: [n]euronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480, 2005.
95. K. J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994.
96. K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.
97. B. G. Galef, Jr and L.-A. Giraldeau. Social influences on foraging in vertebrates: causal mechanisms and adaptive functions. *Animal Behaviour*, 61(1):3–15, 2001.
98. C. Gardiner. *Stochastic Methods*. Springer, Berlin, 2009.
99. C. W. Gardiner et al. *Handbook of Stochastic Methods*, volume 4. Springer Berlin, 1985.
100. M. Gardner. Games - the fantastic combinations of John Conway's new solitaire game life. *Scientific American*, 223:12—123, 1970.
101. C. Gershenson. RBNLab, 2003.
102. C. Gershenson. Introduction to random Boolean networks. In M. Bedau, P. Husbands, T. Hutton, S. Kumar, and H. Suzuki, editors, *Proceedings of the Workshops and Tutorials of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALife IX), Boston, USA*, pages 160–173, 2004.
103. C. Gershenson. Phase transitions in random Boolean networks with different updating schemes, 2004.
104. C. Gershenson. Updating schemes in random Boolean networks: Do they really matter? In J. Pollack, M. Bedau, P. Husbands, T. Ikegami, and R. A. Watson, editors, *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALife IX), Boston, USA*, pages 238–243, Cambridge, USA, 2004. MIT Press.
105. J. Geweke. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.*, 77(378):304–313, 1982.
106. J. Geweke. Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.*, 79(388):907–915, 1984.
107. L.-A. Giraldeau, T. J. Valone, and J. J. Templeton. Potential disadvantages of using socially acquired information. *Proc. Roy. Soc. B*, 357(1427):1559–1566, 2002.
108. R. J. Glauber. Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, 4(2):294–307, 1963.
109. J. K. Goeree and C. A. Holt. Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567, 1999.
110. G. Gómez-Herrero, W. Wu, K. Rutanen, M. Soriano, G. Pipa, and R. Vicente. Assessing coupling dynamics from an ensemble of time series. *Entropy*, 17(4):1958–1970, 2015.
111. B. Gourévitch and J. J. Eggermont. Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97(3):2533–2543, 2007.
112. C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
113. C. W. J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.
114. C. W. J. Granger. Prize lecture: Time series analysis, cointegration, and applications. Nobelprize.org. Nobel Media, December 2003. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2003/granger-lecture.html.
115. C. W. J. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic Press, San Diego, CA, USA, 1977.
116. P. Grassberger. New mechanism for deterministic diffusion. *Physical Review A*, 28(6):3666, 1983.
117. P. Grassberger. Long-range effects in an elementary cellular automaton. *Journal of Statistical Physics*, 45(1-2):27–39, 1986.

118. P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373, 1988.
119. L. Gray. A mathematician looks at Wolfram's new kind of science. *Notices of the American Mathematical Society*, 50(2):200–211, 2003.
120. D. Gregorio, S, R. Serra, and M. Villani. Applying cellular automata in complex environmental problems: The simulation of the bioremediation of contaminated soils. *Theoretical Computer Science*, 217:131–156, 1999.
121. V. Griffith and C. Koch. Quantifying synergistic mutual information. In M. Prokopenko, editor, *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pages 159–190. Springer, Berlin/Heidelberg, 2014.
122. M. M. Hackbart and D. A. Anderson. On measuring economic diversification. *Land Economics*, pages 374–378, 1975.
123. I. Hacking. The logic of Pascal's wager. *American Philosophical Quarterly*, 9(2):186–192, 1972.
124. D. W. Hahs and S. Pethel. Transfer entropy for coupled autoregressive processes. *Entropy*, 15:767–788, 2013.
125. A. G. Haldane and R. M. May. Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355, 2011.
126. J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
127. N. O. Handegard, K. M. Boswell, C. C. Ioannou, S. P. Leblanc, D. B. Tjøstheim, and I. D. Couzin. The dynamics of coordinated group hunting and collective information transfer among schooling prey. *Current Biology*, 22(13):1213–1217, 2012.
128. E. J. Hannan. *Multiple Time Series*. John Wiley, New York, 1970.
129. J. E. Hanson and J. P. Crutchfield. The attractor-basin portait of a cellular automaton. *Journal of Statistical Physics*, 66:1415–1462, 1992.
130. J. E. Hanson and J. P. Crutchfield. Computational mechanics of cellular automata: An example. *Physica D*, 103(1-4):169–189, 1997.
131. M. Harder, C. Salge, and D. Polani. Bivariate measure of redundant information. *Physical Review E*, 87:012130+, 2013.
132. M. Harré and T. Bossomaier. Phase-transition-like behaviour of information measures in financial markets. *EPL (Europhysics Letters)*, 87(1):18009, 2009.
133. M. S. Harré, S. R. Atkinson, and L. Hossain. Simple nonlinear systems and navigating catastrophes. *The European Physical Journal B*, 86(6):1–8, 2013.
134. M. S. Harré and T. Bossomaier. Strategic islands in economic games: Isolating economies from better outcomes. *Entropy*, 16(9):5102–5121, 2014.
135. I. Harvey and T. Bossomaier. Time out of joint: Attractors in asynchronous Boolean networks. In P. Husbands and I. Harvey, editors, *Proceedings of the 4th European Conference on Artificial Life*, pages 67–75, 1997.
136. T. Helvik, K. Lindgren, and M. G. Nordahl. Local information in one-dimensional cellular automata. In P. M. A. Sloot, B. Chopard, and A. G. Hoekstra, editors, *Proceedings of the International Conference on Cellular Automata for Research and Industry, Amsterdam*, volume 3305 of *Lecture Notes in Computer Science*, pages 121–130. Springer, Berlin/Heidelberg, 2004.
137. T. Helvik, K. Lindgren, and M. G. Nordahl. Continuity of information transport in surjective cellular automata. *Communications in Mathematical Physics*, 272(1):53–74, 2007.
138. K. Hlaváčková-Schindler. Equivalence of Granger causality and transfer entropy: A generalization. *Applied Mathematical Sciences*, 5(73):3637–3648, 2011.
139. C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24):10240–10245, 2007.
140. J. Hull. *Options, futures and other derivatives*. Pearson Education, 2009.
141. E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.*, 31:253–258, 1925.
142. S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs. Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. *PLoS ONE*, 6(11):e27431+, 2011.

143. J. Jacod. Multivariate point processes: Predictable projection, Radon-Nikodym derivatives, representation of martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(3):235–253, 1975.
144. M. Jafari-Mamaghani and J. Tyrcha. Transfer entropy expressions for a class of non-Gaussian distributions. *Entropy*, 16:1743–1755, 2014.
145. E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.
146. E. T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, 2003.
147. J. Jost and M. P. Joy. Spectral properties and synchronization in coupled map lattices. *Physical Review E*, 65(1):016201+, 2001.
148. L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.
149. A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166:43–62, 2002.
150. H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, MA, 1997.
151. G. Kastberger, E. Schmelzer, and I. Kranner. Social waves in giant honeybees repel hornets. *PLoS ONE*, 3(9):e3141, 2008.
152. S. Katare and D. H. West. Optimal complex networks spontaneously emerge when information transfer is maximized at least expense: A design perspective. *Complexity*, 11(4):26–35, 2006.
153. Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin. Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences*, 108(46):18720–18725, 2011.
154. S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York, 1993.
155. J. L. Kelly. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.
156. J. M. Keynes. *General theory of employment, interest and money*. Atlantic Books, 2006.
157. A. Kim, D. Putrino, S. Ghosh, and E. N. Brown. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput. Biol.*, 7(3):e1001110, 03 2011.
158. H. Kim, P. Davies, and S. I. Walker. New scaling relation for information transfer in biological networks, 2015. arXiv:1508.04174.
159. J. Kim, G. Kim, S. An, Y.-K. Kwon, and S. Yoon. Entropy-based analysis and bioinformatics-inspired integration of global economic information transfer. *PloS One*, 8(1):e51986, 2013.
160. K. Kiyono, Z. R. Struzik, and Y. Yamamoto. Criticality and phase transition in stock-price fluctuations. *Physical Review Letters*, 96(6):068701, 2006.
161. A. S. Klyubin, D. Polani, and C. L. Nehaniv. All else being equal be empowered. In M. S. Capcarrère, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, editors, *8th European Conference on Artificial Life (ECAL 2005)*, volume 3630 of *Lecture Notes in Computer Science*, pages 744–753. Springer, Berlin / Heidelberg, 2005.
162. A. S. Klyubin, D. Polani, and C. L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12), 2008.
163. A. Kolmogorov. *Limit distributions for sums of independent random variables*. Addison-Wesley, 1954.
164. A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, NY, 1956.
165. G. Korniss. Synchronization in weighted uncorrelated complex networks in a noisy environment: Optimization and connections with transport efficiency. *Physical Review E*, 75(5):051121+, 2007.
166. L. Kozachenko and N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.
167. A. Kraskov. *Synchronization and Interdependence Measures and Their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data*, volume 24 of *Publication Series of the John von Neumann Institute for Computing*. John von Neumann Institute for Computing, Jülich, Germany, 2004.

168. A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138–066153, 2004.
169. Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, volume 39, pages 420–422. Springer Berlin / Heidelberg, 1975.
170. Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag, 1984.
171. O. Kwon and G. Oh. Asymmetric information flow between market index and individual stocks in several stock markets. *EPL (Europhysics Letters)*, 97(2):28007, 2012.
172. R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.
173. R. Landauer. Information is physical. *Physics Today*, 44(23):23–29, 1991.
174. C. G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D*, 42(1-3):12–37, 1990.
175. V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.
176. H. W. Lau and P. Grassberger. Information theoretic aspects of the two-dimensional Ising model. *Phys. Rev. E*, 87(2):022128, 2013.
177. S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson. The metabolic cost of neural information. *Nature Neuroscience*, 1(1):36–41, 1998.
178. N. Levinson. The Wiener RMS (root-mean-square) error criterion in filter design and prediction. *J. Math. Phys.*, 25:261–278, 1947.
179. M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, NY, 1993.
180. M. Lindner, R. Vicente, V. Priesemann, and M. Wibral. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neuroscience*, 12(1):119+, 2011.
181. P. B. S. Lissaman and C. A. Shollenberger. Formation flight of birds. *Science*, 168(3934):1003–1005, 1970.
182. J. T. Lizier. *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer Theses. Springer, Berlin / Heidelberg, 2013.
183. J. T. Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.
184. J. T. Lizier. Measuring the dynamics of information processing on a local scale in time and space. In M. Wibral, R. Vicente, and J. T. Lizier, editors, *Directed Information Measures in Neuroscience*, Understanding Complex Systems, pages 161–193. Springer, Berlin/Heidelberg, 2014.
185. J. T. Lizier, F. M. Atay, and J. Jost. Information storage, loop motifs, and clustered structure in complex networks. *Physical Review E*, 86(2):026110+, 2012.
186. J. T. Lizier, B. Flecker, and P. L. Williams. Towards a synergy-based approach to measuring information modification. In *Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE)*, pages 43–51. IEEE, 2013.
187. J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *Journal of Computational Neuroscience*, 30(1):85–107, 2011.
188. J. T. Lizier and J. R. Mahoney. Moving frames of reference, relativity and invariance in transfer entropy and information dynamics. *Entropy*, 15(1):177–197, 2013.
189. J. T. Lizier, M. Piraveenan, D. Pradhana, M. Prokopenko, and L. S. Yaeger. Functional and structural topologies in evolved neural networks. In G. Kampis, I. Karsai, and E. Szathmáry, editors, *Proceedings of the European Conference on Artificial Life (ECAL), Budapest, Hungary*, volume 5777 of *Lecture Notes in Computer Science*, pages 140–147. Springer, Berlin/Heidelberg, 2011.
190. J. T. Lizier, S. Pritam, and M. Prokopenko. Information dynamics in small-world Boolean networks. *Artificial Life*, 17(4):293–314, 2011.
191. J. T. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *European Physical Journal B*, 73(4):605–615, 2010.

192. J. T. Lizier, M. Prokopenko, I. Tanev, and A. Y. Zomaya. Emergence of glider-like structures in a modular robotic system. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK*, pages 366–373. MIT Press, Cambridge, MA, 2008.
193. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Detecting non-trivial computation in complex dynamics. In Almeida, L. M. Rocha, E. Costa, I. Harvey, and A. Coutinho, editors, *Proceedings of the 9th European Conference on Artificial Life (ECAL 2007)*, volume 4648 of *Lecture Notes in Computer Science*, pages 895–904. Springer, Berlin/Heidelberg, 2007.
194. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. The information dynamics of phase transitions in random Boolean networks. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK*, pages 374–381. MIT Press, Cambridge, MA, 2008.
195. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77(2):026110+, 2008.
196. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Information modification and particle collisions in distributed computation. *Chaos*, 20(3):037109+, 2010.
197. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Coherent information structure in complex computation. *Theory in Biosciences*, 131(3):193–203, Nov. 2012.
198. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local measures of information storage in complex distributed computation. *Information Sciences*, 208:39–54, 2012.
199. J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. A framework for the local information dynamics of distributed computation in complex systems. In M. Prokopenko, editor, *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pages 115–158. Springer Berlin Heidelberg, 2014.
200. J. T. Lizier and M. Rubinov. Multivariate construction of effective computational networks from observational data. Technical Report Preprint 25/2012, Max Planck Institute for Mathematics in the Sciences, 2012.
201. J. T. Lizier and M. Rubinov. Inferring effective computational connectivity using incrementally conditioned multivariate transfer entropy. *BMC Neuroscience*, 14(Suppl 1):P337+, 2013.
202. LoC. National Library of Congress blogs. http://blogs.loc.gov/loc/, Accessed 2014.
203. R. Lowenstein. *When genius failed: The rise and fall of Long-Term Capital Management*. Random House, 2000.
204. Q. Lu and C. Teuscher. Damage spreading in spatial and small-world random Boolean networks. *Physical Review E*, 89(2):022806+, 2014.
205. M. Lungarella, A. Pitti, and Y. Kuniyoshi. Information transfer at multiple scales. *Physical Review E*, 76(5):056117+, 2007.
206. M. Lungarella and O. Sporns. Mapping information flow in sensorimotor networks. *PLoS Computational Biology*, 2(10):e144+, 2006.
207. H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2005.
208. D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.
209. D. MacKenzie. Long-Term Capital Management and the sociology of arbitrage. *Economy and Society*, 32(3):349–380, 2003.
210. M. Madulara, P. Francisco, S. Nawang, D. Arogancia, C. Cellucci, P. Rapp, and A. Albano. EEG transfer entropy tracks changes in information transfer on the onset of vision. *Int. J. Mod. Phys.*, 1(1):1–5, 2010.
211. V. Mäki-Marttunen, I. Diez, J. M. Cortes, D. R. Chialvo, and M. Villarreal. Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Frontiers in Neuroinformatics*, 7:24+, 2013.
212. R. N. Mantegna, H. E. Stanley, et al. *An Introduction to Econophysics: Correlations and Complexity in Finance*, volume 9. Cambridge University Press, Cambridge, 2000.

213. D. Marinazzo, M. Pellicoro, and S. Stramaglia. Causal information approach to partial conditioning in multivariate data sets. *Computational and Mathematical Methods in Medicine*, 2012:303601+, 2012.
214. D. Marinazzo, M. Pellicoro, G. Wu, L. Angelini, J. M. Cortés, and S. Stramaglia. Information transfer and criticality in the Ising model on the human connectome. *PLoS ONE*, 9(4):e93616+, 2014.
215. D. Marinazzo, G. Wu, M. Pellicoro, L. Angelini, and S. Stramaglia. Information flow in networks and the law of diminishing marginal returns: evidence from modeling and human electroencephalographic recordings. *PloS ONE*, 7(9):e45026+, 2012.
216. R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(2):275–281, 2002.
217. G. Martius, R. Der, and N. Ay. Information driven self-organization of complex robotic behaviors. *PLoS ONE*, 8(5):e63400+, 2013.
218. H. Matsuda, K. Kudo, R. Nakamura, O. Yamakawa, and T. Murata. Mutual information of Ising systems. *Int. J. Theor. Phys.*, 35(4):839–845, 1996.
219. R. M. May, S. A. Levin, and G. Sugihara. Complex systems: Ecology for bankers. *Nature*, 451(7181):893–895, 2008.
220. S. S. McMahon, A. Sim, S. Filippi, R. Johnson, J. Liepe, D. Smith, and M. P. H. Stumpf. Information theory and signal transduction systems: From molecular information processing to network inference. *Seminars in Cell & Developmental Biology*, 35:98–108, 2014.
221. A. D. R. McQuarrie and C.-L. Tsai. *Regression and Time Series Model Selection*. World Scientific Publishing, Singapore, 1998.
222. S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
223. G. Miller. Notes on the bias of information estimates. In H. Quastler, editor, *Information in Psychology II-B*, pages 95–100. Free Press, Glencoe, IL, 1955.
224. J. M. Miller, A. Kolpas, J. P. Juchem Neto, and L. F. Rossi. A continuum three-zone model for swarms. *Bulletin of Mathematical Biology*, 74:536–561, 2011.
225. M. Mitchell. Computation in cellular automata: A selected review. In T. Gramss, S. Bornholdt, M. Gross, M. Mitchell, and T. Pellizzari, editors, *Non-Standard Computation*, pages 95–140. VCH, Weinheim, 1998.
226. M. Mitchell. Complex systems: Network thinking. *Artificial Intelligence*, 170(18):1194–1212, 2006.
227. M. Mitchell. *Complexity: A guided tour*. Oxford University Press, New York, 2009.
228. M. Mitchell, J. P. Crutchfield, and P. T. Hraber. Dynamics, computation, and the "edge of chaos": A re-examination. In G. Cowan, D. Pines, and D. Melzner, editors, *Complexity: Metaphors, Models, and Reality*, volume 19 of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 497–513. Addison-Wesley, Reading, MA, 1994.
229. M. Mitchell, J. P. Crutchfield, and P. T. Hraber. Evolving cellular automata to perform computations: mechanisms and impediments. *Physica D*, 75:361–391, 1994.
230. A. Montalto, L. Faes, and D. Marinazzo. MuTE: A MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy. *PLoS ONE*, 9(10):e109462+, 2014.
231. Y.-I. Moon, B. Rajagapolan, and U. Lall. Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, 52(3):2318–2321, 1995.
232. M. Morf, A. Viera, D. T. L. Lee, and T. Kailath. Recursive multichannel maximum entropy spectral estimation. *IEEE Trans. Geosci. Elec.*, 16(2):85 –94, 1978.
233. F. Mosteller and J. W. Tukey. *Data Analysis and Regression. A Second Course in Statistics*. Addison-Wesley, Reading, MA, 1977.
234. K. Nagel and S. Rasmussen. Traffic at the edge of chaos. In R. Brooks and P. Maes, editors, *Proc. ALife IV*, pages 222–235. MIT Press, 1994.
235. K. Nakajima, T. Li, R. Kang, E. Guglielmino, D. G. Caldwell, and R. Pfeifer. Local information transfer in soft robotic arm. In *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1273–1280. IEEE, 2012.

236. G. Nason. A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):879–904, 2013.
237. M. Nawrot, A. Aertsen, and S. Rotter. Single-trial estimation of neuronal firing rates: From single-neuron spike trains to population activity. *J. Neurosci. Meth.*, 94(1):81–92, 1999.
238. A. G. Nedungadi, G. Rangarajan, N. Jain, and M. Ding. Analyzing multiple spike trains with nonparametric Granger causality. *J. Comput. Neurosci.*, 27:55–64, 2009.
239. I. Nemenman, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck. Neural coding of natural stimuli: Information at sub-millisecond resolution. *PLoS Comput. Biol.*, 4(3):e1000025, 03 2008.
240. I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, volume 14. The MIT Press, Cambridge, MA, USA, Sept. 2002.
241. M. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341–346, 1999.
242. J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A:175–240, 1928.
243. J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. A*, 231:289–337, 1933.
244. S. A. Neymotin, K. M. Jacobs, A. A. Fenton, and W. W. Lytton. Synaptic information transfer in computer models of neocortical columns. *Journal of Computational Neuroscience*, 30(1):69–84, 2011.
245. O. Obst, J. Boedecker, and M. Asada. Improving recurrent neural network performance using transfer entropy. In K. Wong, B. Mendis, and A. Bouzerdoum, editors, *Neural Information Processing. Models and Applications*, volume 6444 of *Lecture Notes in Computer Science*, chapter 24, pages 193–200. Springer, Berlin / Heidelberg, 2010.
246. M. Oka and T. Ikegami. Characterizing autonomy in the web via transfer entropy network. In *Proceedings of the Thirteenth International Conference on the Simulation and Synthesis of Living Systems (Artificial Life 13)*, pages 234–242. MIT Press, 2012.
247. M. Oka and T. Ikegami. Exploring default mode and information flow on the web. *PLoS ONE*, 8(4):e60398+, 2013.
248. M. Okatan, M. Wilson, and E. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Computation*, 17(9):1927–61, 2005.
249. J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5):056110, 2003.
250. R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011:156869+, 2011.
251. N. H. Packard. Adaptation toward the edge of chaos. In J. A. S. Kelso, A. J. Mandell, and M. F. Shlesinger, editors, *Dynamic Patterns in Complex Systems*, pages 293–301. World Scientific, Teanek, NJ, 1988.
252. J. Pahle, A. K. Green, C. J. Dixon, and U. Kummer. Information transfer in signalling pathways: A study using coupled simulated and experimental data. *BMC Bioinformatics*, 9:139, 2008.
253. M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová. Synchronization as adjustment of information rates: Detection from bivariate time series. *Phys. Rev. E*, 63(4):046211, 2001.
254. L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.
255. S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7(1):87–107, 1996.
256. A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 1984.
257. J. K. Parrish and L. Edelstein-Keshet. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 284(5411):99–101, 1999.

258. B. L. Partridge. The structure and function of fish schools. *Scientific American*, 246(6):114–123, 1982.
259. R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge, 2004.
260. D. Peak, J. D. West, S. M. Messinger, and K. A. Mott. Evidence for complex, collective dynamics and emergent, distributed computation in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):918–922, 2004.
261. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.
262. A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, Cambridge, 2003.
263. D. Polani. Information: Currency of life? *HFSP Journal*, 3(5):307–316, 2009.
264. D. Polani and M. Möller. Models of information processing in the sensorimotor loop. In F. Emmert-Streib and M. Dehmer, editors, *Information Theory and Statistical Learning*, pages 289–308. Springer, 2009.
265. A. Porta, G. Baselli, F. Lombardi, N. Montano, A. Malliani, and S. Cerutti. Conditional entropy approach for the evaluation of the coupling strength. *Biological Cybernetics*, 81(2):119–129, 1999.
266. W. K. Potts. The chorus-line hypothesis of manoeuvre coordination in avian flocks. *Nature*, 309:344–345, 1984.
267. A. Pregowska, J. Szczepanski, and E. Wajnryb. Mutual information against correlations in binary communication channels. *BMC Neuroscience*, 16(1):32, 2015.
268. M. Prokopenko. Guided self-organization. *HFSP Journal*, 3(5):287–289, 2009.
269. M. Prokopenko, editor. *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*. Springer, Berlin/Heidelberg, 2014.
270. M. Prokopenko, F. Boschietti, and A. J. Ryan. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28, 2009.
271. M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. In S. Nolfi, G. Baldassarre, R. Calabretta, J. C. T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, and D. Parisi, editors, *From Animals to Animats 9: Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06)*, volume 4095 of *Lecture Notes in Computer Science*, pages 558–569. Springer, Berlin Heidelberg, 2006.
272. M. Prokopenko, V. Gerasimov, and I. Tanev. Measuring spatiotemporal coordination in a modular robotic system. In *Proceedings of the 10th International Conference on the Simulation and Synthesis of Living Systems (ALifeX), Bloomington, Indiana, USA*. MIT Press, 2006.
273. M. Prokopenko and J. T. Lizier. Transfer entropy and transient limits of computation. *Scientific Reports*, 4:5394+, 2014.
274. M. Prokopenko, J. T. Lizier, O. Obst, and X. R. Wang. Relating Fisher information to order parameters. *Phys. Rev. E*, 84:041116, Oct 2011.
275. M. Prokopenko, J. T. Lizier, and D. C. Price. On thermodynamic interpretation of transfer entropy. *Entropy*, 15(2):524–543, 2013.
276. R. Quax, A. Apolloni, and P. M. A. Sloot. The diminishing role of hubs in dynamical processes on complex networks. *Journal of The Royal Society Interface*, 10(88):20130568, 2013.
277. C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience*, 30(1):17–44, 2011.
278. D. V. Radakov. *Schooling in the ecology of fish*. John Wiley, New York, 1973. Translated from Russian by H. Mills.
279. M. Ragwitz and H. Kantz. Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Physical Review E*, 65(5):056201+, 2002.
280. P. Rämö, S. Kauffman, J. Kesseli, and O. Yli-Harja. Measures for information propagation in Boolean networks. *Physica D*, 227(1):100–104, 2007.

281. Y. Reddy and A. Sebastin. Research interaction between forex and stock markets in India: An entropy approach. *Vikalpa*, 33(4):27, 2008.
282. C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87 Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, volume 21, pages 25–34, New York, NY, USA, 1987. ACM.
283. A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. E. S. Socolar. Mutual information in random Boolean models of regulatory networks. *Physical Review E*, 77(1):011901–10, 2008.
284. D. R. Rigney, A. L. Goldberger, W. Ocasio, Y. Ichimaru, G. B. Moody, and R. Mark. Multichannel physiological data: Description and analysis. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 105–129. Addison-Wesley, Reading, MA, 1993.
285. J. Rissanen and M. Wax. Measures of mutual and causal dependence between two time series. *IEEE Trans. Inf. Theory*, 33(4):598–601, 1987.
286. J. W. Rivkin. Reproducing knowledge: Replication without imitation at moderate complexity. *Organisation Science*, 12(3):274–293, May-June 2001.
287. J. Rodewald, J. Colombi, K. Oyama, and A. Johnson. Using information-theoretic principles to analyze and evaluate complex adaptive supply network architectures. *Procedia Computer Science*, 61:147–152, 2015.
288. Y. Roudi, S. Nirenberg, and P. E. Latham. Pairwise maximum entropy models for studying large biological systems: When they can work and when they can't. *PLoS Comput. Biol.*, 5(5):e1000380, 05 2009.
289. M. S. Roulston. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(3-4):285–294, 1999.
290. Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.
291. J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical Review Letters*, 108:258701, 2012.
292. K. Rutanen. Tim 1.2.0, 2011. Software, http://www.cs.tut.fi/%7Etimhome/tim-1.2.0/tim.htm.
293. K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.
294. X. San Liang and R. Kleeman. Information transfer between dynamical system components. *Phys. Rev. Lett.*, 95:244101, 2005.
295. A. Sanderson. Adaptive filtering of neuronal spike train data. *IEEE Trans. Biomed. Eng.*, 27:271–274, 1980.
296. L. Sandoval. Structure of a global network of financial companies based on transfer entropy. *Entropy*, 16(8):4443–4482, 2014.
297. M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel, I. A. van de Leemput, S. A. Levin, E. H. van Nes, M. Pascual, and J. Vandermeer. Anticipating critical transitions. *Science*, 338(6105):344–348, 2012.
298. T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461–4, 2000.
299. F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009.
300. J. Seok and K. Y. Seon. Mutual information between discrete variables with many categories using recursive adaptive partitioning. *Scientific Reports*, 5:10981, 2015.
301. A. Seth. Causal connectivity of evolved neural networks during behavior. *Network: Computation in Neural Systems*, 16:35–54, 2005.
302. A. K. Seth, A. B. Barrett, and L. Barnett. Causal density and integrated information as measures of conscious level. *Phil. Trans. R. Soc. A*, 369(1952):3748–3767, 2011.
303. C. R. Shalizi, R. Haslinger, J.-B. Rouquier, K. L. Klinkner, and C. Moore. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Physical Review E*, 73(3):036104, 2006.
304. C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

305. H. Shimazaki and S. Shinomoto. Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.*, 29:171–182, 2010.
306. N. Slonim, G. Atwal, G. Tkacik, and W. Bialek. Information-based clustering. *PNAS*, 102:18297–18302, 2005.
307. R. Solé, S. Manrubia, B. Luque, J. Delgado, and J. Bascompte. Phase transitions and complex systems. *Complexity*, 1(4):13–26, 1996.
308. R. V. Solé. *Phase Transitions*. Princeton University Press, 2011.
309. R. V. Solé and S. Valverde. Information transfer and phase transitions in a model of internet traffic. *Physica A*, 289(3-4):595–605, 2001.
310. V. Sperati, V. Trianni, and S. Nolfi. Evolving coordinated group behaviours through maximisation of mean mutual information. *Swarm Intelligence*, 2(2-4):73–95, 2008.
311. O. Sporns. *Networks of the Brain*. MIT Press, Cambridge, Massachusetts, USA, 2011.
312. O. Sporns and M. Lungarella. Evolving coordinated behavior by maximizing information structure. In L. M. Rocha, L. S. Yaeger, M. A. Bedau, D. Floeano, R. L. Goldstone, and A. Vespignani, editors, *Proceedings of the Tenth International Conference on Simulation and Synthesis of Living Systems (ALifeX), Bloomington, Indiana, USA*, pages 323–329. MIT Press, 2006.
313. M. Staniek and K. Lehnertz. Symbolic transfer entropy. *Physical Review Letters*, 100(15), 2008.
314. O. Stetter, D. Battaglia, J. Soriano, and T. Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comput. Biol.*, 8(8):e1002653+, 2012.
315. S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo. Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6):066211+, 2012.
316. S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo. Expanding the transfer entropy to identify information subgraphs in complex systems. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3668–3671. IEEE, 2012.
317. S. Strogatz. *Sync: The Emerging Science of Spontaneous Order*. Hyperion Books, New York, 2003.
318. S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197–200, 1 1998.
319. G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338:496–500, 2012.
320. F. Takens. Detecting strange attractors in turbulence. In D. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, chapter 21, pages 366–381. Springer, Berlin / Heidelberg, 1981.
321. A. Tang, C. Honey, J. Hobbs, A. Sher, A. Litke, O. Sporns, and J. Beggs. Information flow in local cortical networks is not democratic. *BMC Neuroscience*, 9(Suppl 1), 2008.
322. T. Tassier and F. Menczer. Emerging small-world referral networks in evolutionary labor markets. *IEEE Transactions on Evolutionary Computation*, 5(5):482 –492, 2001.
323. J.-P. Thivierge. Scale-free and economical features of functional connectivity in neuronal networks. *Physical Review E*, 90(2):022721+, 2014.
324. S. Thorpe, A. Delorme, and R. V. Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715–725, 2001.
325. N. Timme, W. Alford, B. Flecker, and J. M. Beggs. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. 36(2):119–140, 2014.
326. G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037, 1994.
327. M. Tribus. An engineer looks at Bayes. In *Proc. 7th Annual Workshop Maximum Entropy and Bayesian Methods*, page 31. Springer, 1987.
328. W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.*, 93:1074–1089, 2005.

329. T. Q. Tung, T. Ryu, K. H. Lee, and D. Lee. Inferring gene regulatory networks from microarray time series data using transfer entropy. In P. Kokol, V. Podgorelec, D. Mičetič-Turk, M. Zorman, and M. Verlič, editors, *Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS '07), Maribor, Slovenia*, pages 383–388, Los Alamitos, USA, 2007. IEEE.
330. G. E. Uhlenbeck and L. S. Ornstein. On the theory of Brownian motion. *Phys. Rev.*, 36:823–841, 1930.
331. V. V. Solo. Likelihood functions for multivariate point processes with coincidences. In *46th IEEE Conference on Decision and Control, 2007*, pages 4245–4250, Dec. 2007.
332. V. A. Vakorin, O. A. Krakovska, and A. R. McIntosh. Confounding effects of indirect connections on causality estimation. *Journal of Neuroscience Methods*, 184(1):152–160, 2009.
333. F. Vanni, M. Luković, and P. Grigolini. Criticality and transmission of information in a swarm of cooperative units. *Physical Review Letters*, 107(7), 2011.
334. G. Ver Steeg and A. Galstyan. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 3–12, New York, NY, USA, 2013. ACM.
335. P. F. Verdes. Assessing causality from multivariate time series. *Physical Review E*, 72(2):026222–9, 2005.
336. R. Vicente and M. Wibral. Efficient estimation of information transfer. In M. Wibral, R. Vicente, and J. T. Lizier, editors, *Directed Information Measures in Neuroscience*, Understanding Complex Systems, pages 37–58. Springer, Berlin/Heidelberg, 2014.
337. R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1):45–67, 2011.
338. T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75:1226–1229, 1995.
339. M. Vinck, F. Battaglia, V. Balakirsky, A. Vinck, and C. Pennartz. Estimation of the entropy based on its polynomial representation. *Phys. Rev. E*, 85:051139, 2012.
340. V. Q. Vu, B. Yu, and R. E. Kass. Information in the nonstationary case. *Neural computation*, 21:688–703, 2009.
341. A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.*, 54(3):426–482, 1943.
342. S. I. Walker, L. Cisneros, and P. C. W. Davies. Evolutionary transitions and top-down causation. In *Artificial Life 13*, pages 283–290. MIT Press, 2012.
343. S. I. Walker, H. Kim, and P. C. W. Davies. The informational architecture of the cell, 2015. arXiv:1507.03877.
344. X. R. Wang, J. M. Miller, J. T. Lizier, M. Prokopenko, and L. F. Rossi. Measuring information storage and transfer in swarms. In T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, and R. Doursat, editors, *Advances in Artificial Life, ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*, pages 838–845. MIT Press, 2011.
345. X. R. Wang, J. M. Miller, J. T. Lizier, M. Prokopenko, and L. F. Rossi. Quantifying and tracing information cascades in swarms. *PLoS ONE*, 7(7):e40084+, 2012.
346. D. J. Watts. *Six Degrees: The Science of a Connected Age*. Norton, New York, 2003.
347. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.
348. M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente. Measuring information-transfer delays. *PLoS ONE*, 8(2):e55809+, 2013.
349. M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser. Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. *Progress in Biophysics and Molecular Biology*, 105(1-2):80–97, 2011.
350. M. Wibral, R. Vicente, and M. Lindner. Transfer entropy in neuroscience. In M. Wibral, R. Vicente, and J. T. Lizier, editors, *Directed Information Measures in Neuroscience*, Understanding Complex Systems, pages 3–36. Springer, Berlin/Heidelberg, 2014.

351. M. Wibral, R. Vicente, and J. T. Lizier, editors. *Directed Information Measures in Neuroscience*. Springer, Berlin, Heidelberg, 2014.
352. M. Wibral, P. Wollstadt, U. Meyer, N. Pampu, V. Priesemann, and R. Vicente. Revisiting Wiener's principle of causality – interaction-delay reconstruction using transfer entropy and multivariate analysis on delay-weighted graphs. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 3676–3679. IEEE, 2012.
353. R. T. Wicks, S. C. Chapman, and R. Dendy. Mutual information as a tool for identifying phase transitions in dynamical complex systems with limited data. *Physical Review E*, 75, 5 2007.
354. N. Wiener. The theory of prediction. In E. F. Beckenback, editor, *Modern Mathematics for Engineers*, 1, chapter 8. McGraw-Hill, New York, 1956.
355. R. A. Wiggins and E. A. Robinson. Recursive solution of the multichannel filtering problem. *J. Geophys. Res.*, 70:1885–1891, 1965.
356. S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 6(1):60–62, 1938.
357. P. L. Williams and R. D. Beer. Information dynamics of evolved agents. In S. Doncieux, B. Girard, A. Guillot, J. Hallam, J.-A. Meyer, and J.-B. Mouret, editors, *From Animals to Animats 11*, volume 6226 of *Lecture Notes in Computer Science*, chapter 4, pages 38–49. Springer, Berlin / Heidelberg, 2010.
358. P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. 2010. arXiv:1004.2515.
359. P. L. Williams and R. D. Beer. Generalized measures of information transfer. 2011. arXiv:1102.1507.
360. J. Wilms, M. Troyer, and F. Verstraete. Mutual information in classical spin models. *J. Stat. Mech.*, P!0011:1–16, 2011.
361. S. Wolfram, editor. *Theory and Applications of Cellular Automata*. World Scientific, 1986.
362. S. Wolfram. *A New Kind of Science*. Wolfram Media, Champaign, IL, USA, 2002.
363. P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, F. J. Díaz-Pernas, and M. Wibral. Efficient transfer entropy analysis of non-stationary neural time series. *PLoS ONE*, 9(7):e102833+, 2014.
364. D. H. Wolpert, M. Harré, E. Olbrich, N. Bertschinger, and J. Jost. Hysteresis effects of changing the parameters of noncooperative games. *Physical Review E*, 85(3):036102, 2012.
365. X. Wu, W. Wang, and W. X. Zheng. Inferring topologies of complex networks with hidden variables. *Physical Review E*, 86:046106+, 2012.
366. A. Wuensche. Classifying cellular automata automatically: Finding gliders, filtering, and relating space-time patterns, attractor basins, and the Z parameter. *Complexity*, 4(3):47–66, 1999.
367. L. Yaeger and O. Sporns. Evolution of neural structure and complexity in a computational ecology. In L. M. Rocha, L. S. Yaeger, M. A. Bedau, D. Floeano, R. L. Goldstone, and A. Vespignani, editors, *Proceedings of the Tenth International Conference on Simulation and Synthesis of Living Systems (ALifeX), Bloomington, Indiana, USA*, pages 330–336. MIT Press, 2006.
368. W. Yoshida, R. J. Dolan, and K. J. Friston. Game theory of mind. *PLoS Computational Biology*, 4(12):e1000254, 2008.
369. X. Zhang and Q. Zhao. Effects of small world topology on the critical boundary for Boolean networks. *Physica A*, 388(17):3657–3666, 2009.
370. D. Zhou, Y. Zhang, Y. Xiao, and D. Cai. Analysis of sampling artifacts on the Granger causality analysis for topology extraction of neuronal dynamics. *Front. Comput. Neurosci.*, 8(75), 2014.
371. D. P. Zitterbart, B. Wienecke, J. P. Butler, and B. Fabry. Coordinated movements prevent jamming in an Emperor penguin huddle. *PLoS ONE*, 6(6), 2011.

Index

T. Bossomaier et al., *An Introduction to Transfer Entropy*,
DOI 10.1007/978-3-319-43222-9