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.

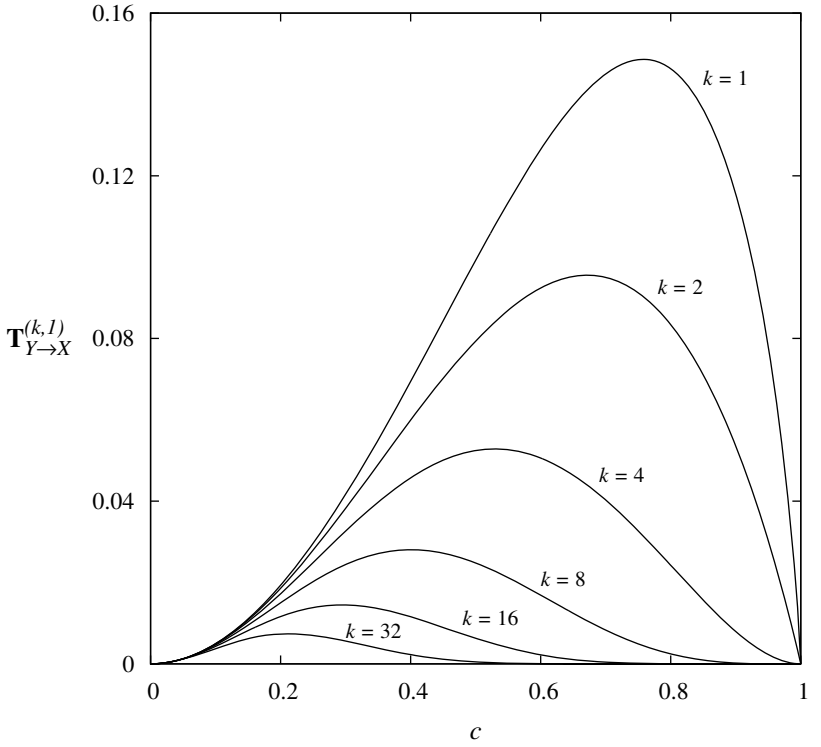

Fig. 4.1 Transfer entropy $\mathbf{T}^{(k,1)}_{Y\to X} = 0$ plotted against coupling parameter c for increasing target history length k for Example 4.1

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.

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

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.

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

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]:

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

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:

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

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

and similarly it is easy to verify that

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

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)

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

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

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

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 $\mathbf{T}_{Y\to X}^{(k,\ell)} \neq 0$. 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.

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

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.