Frontiers in **Robotics and AI**

Methods
16 October 2014

Supplementary Material: JIDT: An information-theoretic toolkit for studying the dynamics of complex systems

Joseph T. Lizier [1,2,*]

[1] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
[2] *CSIRO Digital Productivity Flagship, Marsfield, NSW, Australia*

Correspondence*:
Joseph T. Lizier
CSIRO Digital Productivity Flagship, Corner Vimiera and Pembroke Rds,
Marsfield, NSW, 2122, Australia, joseph.lizier@gmail.com

S.1 INFORMATION-THEORETIC MEASURES

In this section, we give an overview of the information-theoretic measures which are implemented in JIDT. We begin by describing basic information-theoretic measures such as entropy and mutual information in Section S.1.1, then go on to describe in Section S.1.2 the more contemporary measures which are being used to quantify the information dynamics of distributed computation. The latter are the real focus of the toolkit. We also describe in Section S.1.3 how one can measure local or pointwise information-theoretic measures (to assign information values to specific observations or outcomes of variables and their interactions), the extension of the measures to continuous variables in Section S.1.4, and in Section S.1.5 how one can evaluate the statistical significance of the interaction between variables. All features discussed are available in JIDT unless otherwise noted.

S.1.1 BASIC INFORMATION-THEORETIC MEASURES

We first outline basic information-theoretic measures (**Cover and Thomas**, 1991; **MacKay**, 2003) implemented in JIDT.

The fundamental quantity of information theory is the **Shannon entropy**, which represents the expected or average uncertainty associated with any measurement x of a random variable X:[1]

$$H(X) = -\sum_{x \in \alpha_x} p(x) \log_2 p(x). \tag{S.1}$$

with a probabilities distribution function p defined over the alphabet α_x of possible outcomes for x (where $\alpha_x = \{0, \ldots, M_X - 1\}$ without loss of generality for some M_X discrete symbols). Note that unless otherwise stated, logarithms are taken by convention in base 2, giving units in bits.

The Shannon entropy was originally derived following an axiomatic approach, being derived as the unique formulation (up to the base of the logarithm) satisfying a certain set of properties or axioms (see **Shannon** (1948) for further details). The uncertainty $H(X)$ associated with a measurement of X is equal to the expected information required to predict it (see self-information below). $H(X)$ for a measurement

[1] Notation for all quantities is summarised in Table 1.

x of X can also be interpreted as the minimal expected or average number of bits required to encode or describe its value without losing information (**MacKay**, 2003; **Cover and Thomas**, 1991).

The **joint entropy** of two random variables X and Y is a generalization to quantify the uncertainty of their joint distribution:

$$H(X,Y) = -\sum_{x\in\alpha_x}\sum_{y\in\alpha_y} p(x,y)\log_2 p(x,y). \tag{S.2}$$

We can of course write the above equation for multivariate $\mathbf{Z} = \{X, Y\}$, and then generalise to $H(\mathbf{X})$ for $\mathbf{X} = \{X_1, X_2, \ldots, X_G\}$. Such expressions for entropies of multivariates allows us to expand *all* of the following quantities for multivariate $\mathbf{X}$, $\mathbf{Y}$ etc.

The **conditional entropy** of X given Y is the expected uncertainty that remains about x when y is known:

$$H(X \mid Y) = -\sum_{x\in\alpha_x}\sum_{y\in\alpha_y} p(x,y)\log_2 p(x \mid y). \tag{S.3}$$

The conditional entropy for a measurement x of X can be interpreted as the minimal expected number of bits required to encode or describe its value without losing information, given that the receiver of the encoding already knows the value y of Y. The previous quantities are related by the following *chain rule*:

$$H(X,Y) = H(X) + H(Y \mid X). \tag{S.4}$$

The **mutual information** (MI) between X and Y measures the expected reduction in uncertainty about x that results from learning the value of y, or vice versa:

$$I(X;Y) = \sum_{x\in\alpha_x}\sum_{y\in\alpha_y} p(x,y)\log_2 \frac{p(x \mid y)}{p(x)} \tag{S.5}$$

$$= H(X) - H(X \mid Y). \tag{S.6}$$

The MI is symmetric in the variables X and Y. The mutual information for measurements x and y of X and Y can be interpreted as the expected number of bits *saved* in encoding or describing x given that the receiver of the encoding already knows the value of y, in comparison to the encoding of x without the knowledge of y. These descriptions of x with and without the value of y are both minimal without losing information. Note that one can compute the *self-information* $I(X;X) = H(X)$. Finally, one may define a generalization of the MI to a set of more than two variables $\mathbf{X} = \{X_1, X_2, \ldots, X_G\}$, known as the **multi-information** or **integration** (**Tononi et al.**, 1994):

$$I(\mathbf{X}) = I(X_1; X_2; \ldots; X_G)$$

$$= \left(\sum_{g=1}^{G} H(X_g)\right) - H(X_1, X_2, \ldots, X_G). \tag{S.7}$$

Equivalently we can split the set into two parts, $\mathbf{X} = \{\mathbf{Y}, \mathbf{Z}\}$, and express this quantity iteratively in terms of the multi-information of its components individually and the mutual information between those

components:

$$I(\mathbf{X}) = I(\mathbf{Y}) + I(\mathbf{Z}) + I(\mathbf{Y};\mathbf{Z}). \tag{S.8}$$

The **conditional mutual information** between X and Y given Z is the mutual information between X and Y when Z is known:

$$I(X;Y \mid Z) = \sum_{x\in\alpha_x}\sum_{y\in\alpha_y}\sum_{z\in\alpha_z} p(x,y,z)\log_2\frac{p(x \mid y,z)}{p(x \mid z)} \tag{S.9}$$

$$= \sum_{x\in\alpha_x}\sum_{y\in\alpha_y}\sum_{z\in\alpha_z} p(x,y,z)\log_2\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)} \tag{S.10}$$

$$= H(X \mid Z) - H(X \mid Y,Z). \tag{S.11}$$

Note that a conditional MI $I(X;Y \mid Z)$ may be either larger or smaller than the related unconditioned MI $I(X;Y)$ (**MacKay**, 2003). Such conditioning removes redundant information in Y and Z about X, but adds synergistic information which can only be decoded with knowledge of both Y and Z (see further description regarding "partial information decomposition", which refers to attempts to tease these components apart, by: (**Williams and Beer**, 2010; **Harder et al.**, 2013; **Griffith and Koch**, 2014; **Lizier et al.**, 2013; **Bertschinger et al.**, 2013)).

One can consider the MI from two variables Y_1, Y_2 jointly to another variable X, $I(X;Y_1,Y_2)$, and using Eq. (S.4), Eq. (S.6) and Eq. (S.11) decompose this into the information carried by the first variable plus that carried by the second conditioned on the first:

$$I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2 \mid Y_1). \tag{S.12}$$

Of course, this *chain rule* generalises to multivariate $\mathbf{Y}$ of dimension greater than two.

S.1.2 MEASURES OF INFORMATION DYNAMICS

Next, we build on the basic measures of information theory to present measures of the dynamics of information processing. We focus on measures of information in *time-series processes* X of the random variables $\{\ldots X_{n-1}, X_n, X_{n+1}\ldots\}$ with process realisations $\{\ldots x_{n-1}, x_n, x_{n+1}\ldots\}$ for countable time indices n.

We briefly review the framework for *information dynamics* which was recently introduced by **Lizier et al.** (2007, 2008, 2010, 2012, 2014) and **Lizier** (2013, 2014). This framework considers how the information in variable X_{n+1} is related to previous variables, e.g. X_n, of the process or other processes, addressing the fundamental question: *"where does the information in a random variable X_{n+1} in a time series come from?"*. As indicated in Fig. 1, this question is addressed in terms of information from the past of process X (i.e. the information *storage*), information contributed from other source processes Y (i.e. the information *transfer*), and how these sources combine (information *modification*). The goal is to decompose the information in the next observation of X, X_{n+1}, in terms of these information sources.

The **entropy rate** is defined as (**Cover and Thomas**, 1991):

$$H'_{\mu X} = \lim_{n\to\infty}\frac{1}{n}H(X_1, X_2, \ldots, X_n) \tag{S.13}$$

$$= \lim_{n\to\infty}\frac{1}{n}H(\mathbf{X}_n^{(n)}), \tag{S.14}$$

(where the limit exists) where we have used $\mathbf{X}_n^{(k)} = \{X_{n-k+1}, \ldots, X_{n-1}, X_n\}$ to denote the k consecutive variables of X up to and including time step *n*, which has realizations $\mathbf{x}_n^{(k)} = \{x_{n-k+1}, \ldots, x_{n-1}, x_n\}$.

This quantity describes the limiting rate at which the entropy of n consecutive measurements of X grow with n. A related definition for a **(conditional) entropy rate** is given by:[2]

$$H_{\mu X} = \lim_{n\to\infty} H(X_n \mid X_1, X_2, \ldots, X_{n-1}) \tag{S.15}$$

$$= \lim_{n\to\infty} H(X_n \mid \mathbf{X}_{n-1}^{(n-1)}). \tag{S.16}$$

For stationary processes X, the limits for the two quantities $H'_{\mu X}$ and $H_{\mu X}$ exist (i.e. the expected entropy rate converges) and are equal (**Cover and Thomas**, 1991).

For our purposes in considering information dynamics, we are interested in the conditional formulation $H_{\mu X}$, since it explicitly describes how one random variable X_n is related to the previous instances $\mathbf{X}_{n-1}^{(n-1)}$. For practical usage, we are particularly interested in estimation of $H_{\mu X}$ with finite-lengths k, and in estimating it regarding the information at different time indices n. That is to say, we use the notation $H_{\mu X_{n+1}}(k)$ to describe finite-k estimates of the conditional entropy rate in X_{n+1} given $\mathbf{X}_n^{(k)}$:

$$H_{\mu X_{n+1}}(k) = H(X_{n+1} \mid \mathbf{X}_n^{(k)}). \tag{S.17}$$

Assuming stationarity we define:

$$H_{\mu X}(k) = H_{\mu X_{n+1}}(k). \tag{S.18}$$

for any n, and of course letting $k = n$ and joining Eq. (S.16) and Eq. (S.17) we have $\lim_{n\to\infty} H_{\mu X_{n+1}}(k) = H_{\mu X}$.

Next, the **effective measure complexity** (**Grassberger**, 1986) or **excess entropy** (**Crutchfield and Feldman**, 2003) quantifies the total amount of structure or memory in the process X, and is computed in terms of the slowness of the approach of the conditional entropy rate estimates to their limiting value:

$$E_X = \sum_{k=0}^{\infty} \big(H_{\mu X}(k) - H_{\mu X}\big). \tag{S.19}$$

When the process X is stationary we may represent the excess entropy as the mutual information between the semi-infinite past and semi-infinite future of the process:

$$E_X = \lim_{k\to\infty} E_X(k), \tag{S.20}$$

$$E_X(k) = I(\mathbf{X}_n^{(k)}; \mathbf{X}_{n+1}^{(k^+)}), \tag{S.21}$$

where $\mathbf{X}_{n+1}^{(k^+)}$ refers to the next k values $\{X_{n+1}, X_{n+2}, \ldots, X_{n+k}\}$ with realizations $\mathbf{x}_{n+1}^{(k^+)} = \{x_{n+1}, x_{n+2}, \ldots, x_{n+k}\}$, and $E_X(k)$ are finite-k estimates of E_X. This formulation is known as the **predictive information** (**Bialek et al.**, 2001), as it highlights that the excess entropy captures the information in a system's past which can also be found in its future. It is the most appropriate formulation for our purposes, since it provides a clear interpretation as information storage. That is, the excess entropy can be viewed in this formulation as measuring information from the past of the process that is stored – potentially in a distributed fashion in external variables – and is used at some point in the future of the process (**Lizier et al.**, 2012). This contrasts with the statistical complexity (**Crutchfield and Young**, 1989; **Shalizi**, 2001), an upper bound to the excess entropy, which measures *all* information which is

[2] Note that we have reversed the use of the primes in the notation from **Cover and Thomas** (1991), in line with **Crutchfield and Feldman** (2003).

relevant to the prediction of the future of the process states; i.e. the stored information which *may be used* in the future (**Lizier et al.**, 2012).

In contrast again, the **active information storage** (AIS) was introduced by **Lizier et al.** (2012) to measure how much of the information from the past of the process X is observed to be *in use* in computing its *next observation*. This measure of information storage more directly addresses our key question of determining the sources of the information in the next observation X_{n+1}. The active information storage is the expected mutual information between realizations $\mathbf{x}_n^{(k)}$ of the past state $\mathbf{X}_n^{(k)}$ (as $k \to \infty$) and the corresponding realizations x_{n+1} of the next value X_{n+1} of process X:

$$A_X = \lim_{k \to \infty} A_X(k), \tag{S.22}$$

$$A_X(k) = I(\mathbf{X}_n^{(k)}; X_{n+1}). \tag{S.23}$$

We note that $\mathbf{x}_n^{(k)}$ are Takens' *embedding vectors* (**Takens**, 1981) with *embedding dimension* k, which capture the underlying *state* of the process X for Markov processes of order k.[3] As such, one needs to at least take k at the Markovian order of X in order to capture all relevant information in the past of X, otherwise (for non-Markovian processes) the limit $k \to \infty$ is theoretically required in general (**Lizier et al.**, 2012). We also note that since:

$$A_X = H(X) - H_{\mu X}, \tag{S.24}$$

then the limit in Eq. (S.22) exists for stationary processes (i.e. $A(X)$ converges with $k \to \infty$) (**Lizier et al.**, 2012).

Arguably the most important measure in this toolkit is the **transfer entropy** (TE) from **Schreiber** (2000). TE captures the concept of information transfer, as the amount of information that a source process provides about a destination (or target) process' next state in the context of the destination's past. Quantitatively, this is the expected mutual information from realizations $\mathbf{y}_n^{(l)}$ of the state $\mathbf{Y}_n^{(l)}$ of a source process Y to the corresponding realizations x_{n+1} of the next value X_{n+1} of the destination process X, conditioned on realizations $\mathbf{x}_n^{(k)}$ of its previous state $\mathbf{X}_n^{(k)}$:

$$T_{Y \to X}(l) = \lim_{k \to \infty} T_{Y \to X}(k, l), \tag{S.25}$$

$$T_{Y \to X}(k, l) = I(\mathbf{Y}_n^{(l)}; X_{n+1} \mid \mathbf{X}_n^{(k)}). \tag{S.26}$$

TE has become a very popular tool in complex systems in general, e.g. (**Williams and Beer**, 2011; **Lungarella and Sporns**, 2006; **Obst et al.**, 2010; **Barnett and Bossomaier**, 2012; **Lizier et al.**, 2008, 2011c; **Boedecker et al.**, 2012), and in computational neuroscience in particular, e.g. (**Vicente et al.**, 2011; **Lindner et al.**, 2011; **Ito et al.**, 2011; **Stramaglia et al.**, 2012; **Lizier et al.**, 2011b). For multivariate Gaussians, the TE is equivalent (up to a factor of 2) to the **Granger causality** (**Barnett et al.**, 2009a).

There are a number of important considerations regarding the use of this measure (see further discussion by **Lizier et al.** (2008); **Lizier** (2014); **Wibral et al.** (2014b,a); and **Vicente and Wibral** (2014)). First, for the embedding vectors $\mathbf{x}_n^{(k)}$ one needs to at least take k larger than the Markovian order of X in order to eliminate any AIS from being redundantly measured in the TE.[4] Then, one may need to extend k to capture synergies generated in x_{n+1} between the source $\mathbf{y}_n^{(l)}$ and earlier values in X. For non-Markovian processes X (or non-Markovian processes when considered jointly with the source), one

[3] We can use an embedding delay τ to give $\mathbf{x}_n^{(k)} = \{x_{n-(k-1)\tau}, \ldots, x_{n-\tau}, x_n\}$, where this helps to better empirically capture the state from a finite sample size. Non-uniform embeddings (i.e. with irregular delays) may also be useful (**Faes et al.**, 2011) (not implemented in JIDT at this stage).

[4] The destination's embedding dimension should be increased before that of the source, for this same reason.

should theoretically take the limit as $k \to \infty$ (**Lizier et al.**, 2008). Setting k in this manner gives the perspective to separate information storage and transfer in the distributed computation in process X, and allows one to interpret the transfer entropy as properly representing information transfer (**Lizier et al.**, 2008; **Lizier and Prokopenko**, 2010).

Also, note that the transfer entropy can be defined for an arbitrary source-destination delay u (**Wibral et al.**, 2013):

$$T_{Y \to X}(k, l, u) = I(\mathbf{Y}^{(l)}_{n+1-u}; X_{n+1} \mid \mathbf{X}^{(k)}_n), \tag{S.27}$$

and indeed that this should be done for the appropriate causal delay $u > 0$. For ease of presentation here, we describe the measures for $u = 1$ only, though all are straightforward to generalise and are implemented with generic u in JIDT.

While the simple setting $l = 1$ is often used, this is only completely appropriate where y_n is directly causal to x_{n+1} and where it is the only direct causal source in Y (**Lizier et al.**, 2008; **Lizier and Prokopenko**, 2010) (e.g. in cellular automata). In general circumstances, one should use an embedded source *state* $\mathbf{y}^{(l)}_n$ (with $l > 1$), in particular where the observations y mask a hidden Markov process that is causal to X (e.g. in brain imaging data), or where multiple past values of Y in addition to y_n are causal to x_{n+1}.

Finally, for proper interpretation as information transfer, Y is constrained among the causal information contributors to X (**Lizier and Prokopenko**, 2010). With that said, the concepts of information transfer and causality are complementary but distinct, and TE should not be thought of as measuring causal effect (**Ay and Polani**, 2008; **Lizier and Prokopenko**, 2010; **Chicharro and Ledberg**, 2012). **Prokopenko et al.** (2013) and **Prokopenko and Lizier** (2014) have also provided a thermodynamic interpretation of transfer entropy, as being proportional to external entropy production, possibly due to irreversibility.

Now, the transfer entropy may also be conditioned on other possible sources Z to account for their effects on the destination. The **conditional transfer entropy**[5] was introduced for this purpose (**Lizier et al.**, 2008, 2010):

$$T_{Y \to X|Z}(l) = \lim_{k \to \infty} T_{Y \to X|Z}(k, l), \tag{S.28}$$

$$T_{Y \to X|Z}(k, l) = I(\mathbf{Y}^{(l)}_n; X_{n+1} \mid \mathbf{X}^{(k)}_n, Z_n), \tag{S.29}$$

Note that Z_n may represent an embedded state of another variable, or be explicitly multivariate. Also, for simplicity Eq. (S.29) does not explicitly show arbitrary delays in the style of Eq. (S.27) for source-destination and conditional-destination relationships, though these may naturally be defined and are implemented in JIDT. Transfer entropies conditioned on other variables have been used in several biophysical and neuroscience applications, e.g. (**Faes et al.**, 2011, 2012; **Stramaglia et al.**, 2012; **Vakorin et al.**, 2009). We typically describe TE measurements which are not conditioned on any other variables (as in Eq. (S.25)) as **pairwise** or **apparent transfer entropy**, and measurements conditioned on *all* other causal contributors to X_{n+1} as **complete transfer entropy** (**Lizier et al.**, 2008). Further, one can consider multivariate sources $\mathbf{Y}$, in which case we refer to the measure $T_{\mathbf{Y} \to X}(k, l)$ as a **collective transfer entropy** (**Lizier et al.**, 2010).

Finally, while how to measure information modification remains an open problem (see **Lizier et al.** (2013)), JIDT contains an implementation of an early attempt at capturing this concept in the **separable**

[5] This is sometimes known as "multivariate" TE, though this term can be confused with TE applied to multivariate source and destination variables (i.e. the collective TE).

information (**Lizier et al.**, 2010):

$$S_X = \lim_{k\to\infty} S_X(k), \tag{S.30}$$

$$S_X(k) = A_X(k) + \sum_{Y\in\mathbf{V}_X\setminus X} T_{Y\to X}(k, l_Y). \tag{S.31}$$

Here, $\mathbf{V}_X$ represents the set of causal information sources $\mathbf{V}_X$ to X, while l_Y is the embedding dimension for source Y.

S.1.3 LOCAL INFORMATION-THEORETIC MEASURES

Local information-theoretic measures (also known as **pointwise information-theoretic measures**) characterise the information attributed with *specific* measurements x, y and z of variables X, Y and Z (**Lizier**, 2014), rather than the traditional expected or average information measures associated with these variables introduced in Section S.1.1 and Section S.1.2. Although they are deeply ingrained in the fabric of information theory, and heavily used in some areas (e.g. in natural language processing (**Manning and Schütze**, 1999)), until recently (**Shalizi**, 2001; **Shalizi et al.**, 2006; **Helvik et al.**, 2004; **Lizier et al.**, 2007, 2008, 2012, 2010) local information-theoretic measures were rarely applied to complex systems.

That these local measures are now being applied to complex systems is important, because they provide a direct, model-free, mechanism to analyse the *dynamics* of how information processing unfolds in time. In other words: traditional (expected) information-theoretic measures would return one value to characterise, for example, the transfer entropy between Y and X. Local transfer entropy on the other hand, returns a time-series of values to characterise the information transfer from Y to X as a function of time, so as to directly reveal the *dynamics* of their interaction. Indeed, it is well-known that local values (within a global average) provide important insights into the dynamics of nonlinear systems (**Dasan et al.**, 2002).

A more complete description of local information-theoretic measurements is provided by **Lizier** (2014). Here we provide a brief overview of the local values of the measures previously introduced.

The most illustrative local measure is of course the **local entropy** or **Shannon information content**. The Shannon information content of an outcome x of measurement of the variable X is (**MacKay**, 2003):

$$h(x) = -\log_2 p(x). \tag{S.32}$$

Note that by convention we use lower-case symbols to denote local information-theoretic measures. The Shannon information content was shown to be the unique formulation for a local entropy (up to the base of the logarithm) satisfying required properties corresponding to those of the expected Shannon entropy (see **Ash** (1965) for details). Now, the quantity $h(x)$ is simply the information content attributed to the specific symbol x, or the information required to predict or uniquely specify that specific value. Less probable outcomes x have higher information content than more probable outcomes, and we have $h(x) \geq 0$. The Shannon information content of a given symbol x is the *code-length* for that symbol in an optimal encoding scheme for the measurements X, i.e. one that produces the minimal expected code length.

We can form all traditional information-theoretic measures as the *average* or *expectation value* of their corresponding local measure, e.g.:

$$H(X) = \sum_{x\in\alpha_x} p(x)h(x), \tag{S.33}$$

$$= \langle h(x)\rangle . \tag{S.34}$$

While the above represents this as an expectation over the relevant ensemble, we can write the same average over all of the N samples x_n (with each sample given an index n) used to generate the probability

distribution function (PDF) $p(x)$ (**Lizier**, 2014; **Lizier et al.**, 2008), e.g.:

$$H(X) = \frac{1}{N}\sum_{n=1}^{N} h(x_n), \tag{S.35}$$

$$= \langle h(x_n) \rangle_n . \tag{S.36}$$

Next, we have the **conditional Shannon information content** (or **local conditional entropy**) (**MacKay**, 2003):

$$h(x \mid y) = -\log_2 p(x \mid y), \tag{S.37}$$

$$h(x, y) = -\log_2 p(x, y), \tag{S.38}$$

$$= h(y) + h(x \mid y), \tag{S.39}$$

$$H(X \mid Y) = \langle h(x \mid y) \rangle . \tag{S.40}$$

As above, local quantities satisfy corresponding chain rules to those of their expected quantities.

The **local mutual information** is defined (uniquely, see **Fano** (1961, ch. 2)) as "the amount of information provided by the occurrence of the event represented by y_i about the occurrence of the event represented by x_i", i.e.:

$$i(x; y) = \log_2 \frac{p(x \mid y)}{p(x)}, \tag{S.41}$$

$$= h(x) - h(x \mid y), \tag{S.42}$$

$$I(X; Y) = \langle i(x; y) \rangle . \tag{S.43}$$

$i(x; y)$ is symmetric in x and y, as is the case for $I(x; y)$. The local mutual information is the difference in code lengths between coding the value x in isolation (under the optimal encoding scheme for X), or coding the value x given y (under the optimal encoding scheme for X given Y). In other words, this quantity captures the coding "cost" for x in not being aware of the value y.

Of course this "cost" averages to be non-negative, however the local mutual information may be either positive or negative for a specific pair x, y. Positive values are fairly intuitive to understand: $i(x; y)$ is positive where $p(x \mid y) > p(x)$, i.e. knowing the value of y *increased* our expectation of (or positively informed us about) the value of the measurement x. Negative values simply occur in Eq. (S.41) where $p(x \mid y) < p(x)$. That is, knowing the value of y changed our belief $p(x)$ about the probability of occurrence of the outcome 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 were x nevertheless occurred. Consider the following example from **Lizier** (2014), of the probability that it will rain today, $p(\texttt{rain} = 1)$, and the probability that it will rain given that the weather forecast said it would not, $p(\texttt{rain} = 1 \mid \texttt{rain\_forecast} = 0)$. We could have $p(\texttt{rain} = 1 \mid \texttt{rain\_forecast} = 0) < p(\texttt{rain} = 1)$, so we would have $i(\texttt{rain} = 1; \texttt{rain\_forecast} = 0) < 0$, because we considered it less likely that rain would occur today when hearing the forecast than without the forecast, in a case where rain nevertheless occurred. Such negative values of MI are actually quite meaningful, and can be interpreted as there being negative information in the value of y about x. We could also interpret the value y as being *misleading* or *misinformative* about the value of x, because it *lowered* our expectation of observing x prior to that observation being made in this instance. In the above example, the weather forecast was misinformative about the rain today.

Note that the local mutual information $i(x; y)$ measure above is distinct from *partial* localization expressions, i.e. the partial mutual information or specific information $I(x; Y)$ (**DeWeese and Meister**, 1999), which consider information contained in specific values x of one variable X about the other

(unknown) variable Y. While there are two valid approaches to measuring partial mutual information, as above there is only one valid approach for the fully local mutual information $i(x;y)$ (**Fano**, 1961, ch. 2).

The **local conditional mutual information** is similarly defined by **Fano** (1961, ch. 2):

$$i(x;y \mid z) = \log_2 \frac{p(x \mid y,z)}{p(x \mid z)}, \tag{S.44}$$

$$= h(x \mid z) - h(x \mid y,z), \tag{S.45}$$

$$I(X;Y \mid Z) = \langle i(x;y \mid z) \rangle . \tag{S.46}$$

$I(X;Y \mid Z)$ is the difference in code lengths (or coding cost) between coding the value x given z (under the optimal encoding scheme for X given Z), or coding the value x given both y and z (under the optimal encoding scheme for X given Y and Z). As per $I(X;Y)$, the local conditional MI is symmetric in x and y, and may take positive or negative values.

The **local multi-information** follows for the observations $x_1, x_2, \ldots, x_G$ as:

$$i(x_1; x_2; \ldots; x_G) = \left(\sum_{g=1}^{G} h(x_g) \right) - h(x_1, x_2, \ldots, x_G). \tag{S.47}$$

Local measures of information dynamics are formed via the local definitions of the basic information-theoretic measures above. Here, the local measures pertain to realisations x_n, $\mathbf{x}_n^{(k)}$, $\mathbf{y}_n^{(l)}$, etc, of the processes at specific time index n. The PDFs may be estimated either from multiple realisations of the process for time index n, or from multiple observations over time from one (or several) full time-series realisation(s) where the process is stationary (see comments by **Lizier** (2014)).

We have the **local entropy rate**:[6]

$$h_{\mu X}(n+1,k) = h(x_{n+1} \mid \mathbf{x}_n^{(k)}), \tag{S.48}$$

$$H_{\mu X}(k) = \left\langle h_{\mu X}(n,k) \right\rangle . \tag{S.49}$$

Next, the **local excess entropy** is defined as (via the predictive information formulation from Eq. (S.21)) (**Shalizi**, 2001):

$$e_X(n+1,k) = i(\mathbf{x}_n^{(k)}; \mathbf{x}_{n+1}^{(k^+)}), \tag{S.50}$$

$$E_X(k) = \langle e_X(n,k) \rangle . \tag{S.51}$$

We then have the **local active information storage** $a_X(n+1)$ (**Lizier et al.**, 2012):

$$a_X(n+1,k) = i(\mathbf{x}_n^{(k)}; x_{n+1}), \tag{S.52}$$

$$A_X(k) = \langle a_X(n+1,k) \rangle . \tag{S.53}$$

The local values of active information storage measure the dynamics of information storage at different time points within a system, revealing to us how the use of memory fluctuates during a process. As

[6] For the local measures of information dynamics, while formal definitions may be provided by taking the limit as $k \to \infty$, we will state only the formulae for their finite-k estimates.

described for the local MI, $a_X(n+1,k)$ may be positive or negative, meaning the past history of the process can either positively inform us or actually *misinform* us about its next value (**Lizier et al.**, 2012). Fig. 1 indicates a local active information storage measurement for time-series process X.

The **local transfer entropy** is (**Lizier et al.**, 2008) (with adjustment for source-destination lag u (**Wibral et al.**, 2013)):

$$t_{Y\to X}(n+1,k,l,u) = i(\mathbf{y}^{(l)}_{n+1-u}; x_{n+1} \mid \mathbf{x}^{(k)}_n), \tag{S.54}$$

$$T_{Y\to X}k,l = \langle t_{Y\to X}(n+1,k,l)\rangle\,. \tag{S.55}$$

These local information transfer values measure the dynamics of transfer in time between a given pair of time-series processes, revealing to us how information is transferred in time and space. Fig. 1 indicates a local transfer entropy measurement for a pair of processes $Y \to X$.

Finally, we have the **local conditional transfer entropy** (**Lizier et al.**, 2008, 2010) (again dropping arbitrary lags and embedding of the conditional here for convenience):

$$t_{Y\to X|Z}(n+1,k,l) = i(\mathbf{y}^{(l)}_n; x_{n+1} \mid \mathbf{x}^{(k)}_n, z_n), \tag{S.56}$$

$$T_{Y\to X|Z}(n+1,k,l) = \left\langle t_{Y\to X|Z}(n+1,k,l)\right\rangle. \tag{S.57}$$

S.1.4 DIFFERENTIAL ENTROPY

Note that all of the information-theoretic measures above considered a discrete alphabet of symbols α_x for a given variable X. When X in fact is a continuous-valued variable, we shift to consider **differential entropy** measurements; see **Cover and Thomas** (1991, ch. 9). We briefly discuss differential entropy, since some of our estimators discussed in Section S.2 evaluate these quantities for continuous-valued variables rather than strictly Shannon entropies.

The differential entropy of a continuous variable X with probability density function $f(x)$ is defined as (**Cover and Thomas**, 1991, ch. 9):

$$H_D(X) = -\int_{S_X} f(x)\log f(x)dx, \tag{S.58}$$

where S_X is the set where $f(x) > 0$. The differential entropy is strongly related to the Shannon entropy, but has important differences to what the Shannon entropy would return on discretizing the same variables. Primary amongst these differences is that $H_D(X)$ changes with scaling of the variable X, and that it can be negative.

Joint and conditional ($H_D(X \mid Y)$) differential entropies may be evaluated from Eq. (S.58) expressions using the same chain rules from the Shannon measures. Similarly, the differential mutual information may be defined as (**Cover and Thomas**, 1991, ch. 9):

$$I_D(X;Y) = H_D(X) - H_D(X \mid Y), \tag{S.59}$$

$$= \int_{S_X,S_Y} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}\,dx\,dy. \tag{S.60}$$

Crucially, the properties of $I_D(X;Y)$ are the same as for discrete variables, and indeed $I_D(X;Y)$ is equal to the discrete MI $I(X^\Delta;Y^\Delta)$ for discretizations X^Δ and Y^Δ with bin size Δ, in the limit $\Delta \to 0$ (**Cover and Thomas**, 1991, ch. 9). Conditional MI and other derived measures (e.g. transfer entropy) follow.

S.1.5 STATISTICAL SIGNIFICANCE TESTING

In *theory*, the MI between two unrelated variables Y and X is equal to 0. The same goes for the TE between two variables Y and X with no directed relationship, or the conditional MI between Y and X given Z where there is no conditional relationship. In *practice*, where the MI, conditional MI or TE are empirically measured from a finite number of samples N, a bias of a non-zero measurement is likely to result even where there is no such (directed) relationship. A common question is then whether a given empirical measurement is statistically different from 0, and therefore represents sufficient evidence for a (directed) relationship between the variables.

This question is addressed in the following manner (**Chávez et al.**, 2003; **Verdes**, 2005; **Vicente et al.**, 2011; **Lindner et al.**, 2011; **Lizier et al.**, 2011b; **Wibral et al.**, 2014a; **Barnett and Bossomaier**, 2012). We form a *null hypothesis* H_0 that there is no such relationship, and then make a test of statistical significance 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 accept the alternate hypothesis that there is a (directed) relationship.

For example, for an MI measurement $I(Y;X)$, we generate the distribution of *surrogate* measurements $I(Y^s;X)$ 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 correlation with X is destroyed. Specifically, this means that $p(x \mid y)$ in Eq. (S.6) is distributed as $p(x)$ (with $p(y)$ retained also).

In some situations, we can compute the distribution of $I(Y^s;X)$ analytically. For example, for linearly-coupled Gaussian multivariates $\mathbf{X}$ and $\mathbf{Y}$, $I(\mathbf{Y}^s;\mathbf{X})$ measured in *nats* follows a chi-square distribution, specifically $\chi^2_{|\mathbf{X}||\mathbf{Y}|}/2N$ with $|\mathbf{X}||\mathbf{Y}|$ degrees of freedom, where $|\mathbf{X}|$ ($|\mathbf{Y}|$) is the number of Gaussian variables in vector $\mathbf{X}$ ($\mathbf{Y}$) (**Geweke**, 1982; **Brillinger**, 2004). Also, for discrete variables X and Y with alphabet sizes M_X and M_Y, $I(Y^s;X)$ measured in *bits* follows a chi-square distribution, specifically $\chi^2_{(M_X-1)(M_Y-1)}/(2N\log 2)$ (**Brillinger**, 2004; **Cheng et al.**, 2006). Note that these distributions are followed *asymptotically* with the number of samples N, and the approach is much slower for discrete variables with skewed distributions (**Barnett**, 2013), *which reduces the utility of this analytic result in practice*.[7] **Barnett and Bossomaier** (2012) generalise these results to state that a model-based null distribution (in *nats*) will follow $\chi^2_d/2N$, where d is the "difference between the number of parameters" in a full model (capturing $p(x \mid y)$ in Eq. (S.6)) and a null model (capturing $p(x)$ only).

Where no analytic distribution is known, the distribution of $I(Y^s;X)$ must be computed empirically. This is done by a resampling method (i.e. permutation or bootstrapping)[8] (**Chávez et al.**, 2003; **Verdes**, 2005; **Vicente et al.**, 2011; **Lindner et al.**, 2011; **Lizier et al.**, 2011b; **Wibral et al.**, 2014a), creating a large number of surrogate time-series pairs $\{Y^s, X\}$ by shuffling (for permutations, or redrawing for bootstrapping) the samples of Y (so as to retain $p(x)$ and $p(y)$ but not $p(x \mid y)$), and computing a population of $I(Y^s;X)$ values.

Now, for a conditional MI, we generate the distribution of $I(Y^s;X \mid Z)$ under H_0, which means that $p(x \mid y,z)$ in Eq. (S.9) is distributed as $p(x \mid z)$ (with $p(y)$ retained also).[9] The asymptotic distribution may be formed analytically for linearly-coupled Gaussian multivariates defined above (**Geweke**, 1982; **Barnett and Bossomaier**, 2012) (in *nats*) as $\chi^2_{|\mathbf{X}||\mathbf{Y}|}/2N$ with $|\mathbf{X}||\mathbf{Y}|$ degrees of freedom – interestingly, this does *not* depend on the Z variable. Similarly, for discrete variables the asymptotic distribution (in

[7] See Section 4.6 for an investigation of this.

[8] JIDT employs permutation tests for resampling.

[9] Clearly, this approach specifically makes a *directional* hypothesis test of Eq. (S.9) rather than a non-directional test of Eq. (S.10). Asymptotically these will be the same anyway (as is clear for the analytic cases discussed here). In practice, we favour this somewhat directional approach since in most cases we are indeed interested in the directional question of whether Y adds information to X in the context of Z.

bits) is $\chi^2_{(M_X-1)(M_Y-1)M_Z}/(2N\log 2)$ (**Cheng et al.**, 2006). Again, the distribution of $I(Y^s; X \mid Z)$ is otherwise computed by permutation (or bootstrapping), this time by creating surrogate time-series $\{Y^s, X, Z\}$ by shuffling (or redrawing) the samples of Y (retaining $p(x \mid z)$ and $p(y)$ but not $p(x \mid y, z)$), and computing a population of $I(Y^s; X \mid Z)$ values.

Statistical significance testing for the transfer entropy can be handled as a special case of the conditional MI. For linear-coupled Gaussian multivariates $\mathbf{X}$ and $\mathbf{Y}$, the null $T_{Y^s \to X}(k, l)$ (in *nats*) is asymptotically $\chi^2/2N$ distributed with $l|\mathbf{X}||\mathbf{Y}|$ degrees of freedom (**Geweke**, 1982; **Barnett and Bossomaier**, 2012; **Barnett**, 2013), while for discrete X and Y, $T_{Y^s \to X}(k, l)$ (in *bits*) is asymptotically $\chi^2/(2N\log 2)$ distributed with $(M_X - 1)(M_Y^l - 1)M_X^k$ degrees of freedom (**Barnett and Bossomaier**, 2012). Again, the distribution of $T_{Y^s \to X}(k, l)$ is otherwise computed by permutation (or bootstrapping) (**Chávez et al.**, 2003; **Verdes**, 2005; **Vicente et al.**, 2011; **Lindner et al.**, 2011; **Lizier et al.**, 2011b; **Wibral et al.**, 2014a), under which surrogates must preserve $p(x_{n+1} \mid x_n^{(k)})$ but not $p(x_{n+1} \mid \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)})$. Directly shuffling the series Y to create the Y^s is *not* a valid approach, since it destroys $\mathbf{y}_n^{(l)}$ vectors (unless $l = 1$). Valid approaches include: shuffling (or redrawing) the $\mathbf{y}_n^{(l)}$ amongst the set of $\{x_{n+1}, \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)}\}$ tuples;[10] rotating the Y time-series (where we have stationarity); or swapping sample time series Y_i between different trials i in an ensemble approach (**Vicente et al.**, 2011; **Wibral et al.**, 2014a; **Lindner et al.**, 2011; **Wollstadt et al.**, 2014). Conditional TE may be handled similarly as a special case of a conditional MI.

Finally, we note that such assessment of statistical significance is often used in the application of effective network inference from multivariate time-series data; e.g. (**Vicente et al.**, 2011; **Lindner et al.**, 2011; **Lizier et al.**, 2011b; **Wibral et al.**, 2014a). In this and other situations where multiple hypothesis tests are considered together, one should correct for multiple comparisons using family-wise error rates (e.g. Bonferroni correction) or false discovery rates.

S.2 ESTIMATION TECHNIQUES

While the mathematical formulation of the quantities in Section S.1 are relatively straightforward, empirically estimating them in practice from a finite number N of samples of time-series data can be a complex process, and is dependent on the type of data you have and its properties. Estimators are typically subject to bias and variance due to finite sample size.

In this section, we introduce the various types of estimators which are included in JIDT. Such estimators are discussed in some depth by **Vicente and Wibral** (2014), for the transfer entropy in particular. Unless otherwise noted, all quoted features and time-complexities are as implemented in JIDT.

S.2.1 DISCRETE-VALUED VARIABLES

For discrete variables X, Y, Z etc, the definitions in Section S.1 may be used directly, by counting the matching configurations in the available data to obtain the relevant plug-in probability estimates (e.g. $\hat{p}(x \mid y)$ and $\hat{p}(x)$ for MI). This approach may be taken for both local and average measures. These estimators are simple and fast, being implemented in $\mathrm{O}(N)$ time even for measures such as transfer entropy which require embedded past vectors (since these may be cached and updated at each step in a time-series). Several bias correction techniques are available, e.g. (**Paninski**, 2003; **Bonachela et al.**, 2008), though not yet implemented in JIDT.

[10] This is the approach taken in JIDT.

S.2.2 CONTINUOUS-VALUED VARIABLES

For continuous variables X, Y, Z, one could simply discretise or bin the data and apply the discrete estimators above. This is a simple and fast approach ($\mathrm{O}(N)$ as above), though it is likely to sacrifice accuracy. Alternatively, we can use an estimator that harnesses the continuous nature of the variables, dealing with the differential entropy and probability density functions. The latter is more complicated but yields a more accurate result. We discuss several such estimators in the following. Note that except where otherwise noted, JIDT implements the most efficient available algorithm for each estimator.

S.2.2.1 Gaussian-distribution model The simplest estimator uses a *multivariate Gaussian model* for the relevant variables, assuming linear interactions between them. Under this model, for $\mathbf{X}$ (of d dimensions) the entropy has the form (**Cover and Thomas**, 1991):

$$H(\mathbf{X}) = \frac{1}{2} \ln \left((2\pi e)^d \mid \Omega_{\mathbf{X}} \mid \right), \tag{S.61}$$

(in *nats*) where $\mid \Omega_{\mathbf{X}} \mid$ is the determinant of the $d \times d$ covariance matrix $\Omega_{\mathbf{X}} = \overline{\mathbf{X}\mathbf{X}^T}$, and the overbar "represents an average over the statistical ensemble" (**Barnett et al.**, 2009b). Any standard information-theoretic measure in Section S.1 can then be obtained from sums and differences of these joint entropies. For example, **Kaiser and Schreiber** (2002) demonstrated how to compute transfer entropy in this fashion. These estimators are fast ($\mathrm{O}\left(Nd^2\right)$) and parameter-free, but subject to the linear-model assumption.

Since PDFs were effectively bypassed in Eq. (S.61), the local entropies (and by sums and differences, other local measures) can be obtained by first reconstructing the probability of a given observation $\mathbf{x}$ in a multivariate process with covariance matrix $\Omega_{\mathbf{X}}$:

$$p(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^d \mid \Omega_{\mathbf{X}} \mid^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Omega_{\mathbf{X}}^{-1} (\mathbf{x} - \mu) \right), \tag{S.62}$$

(where μ is the vector of expectation values of $\mathbf{x}$), then using these values directly in the equation for the given local quantity as a plug-in estimate (**Lizier**, 2014).[11]

S.2.2.2 Kernel estimation Using *kernel-estimators* (e.g. see **Schreiber** (2000) and **Kantz and Schreiber** (1997)), the relevant joint PDFs (e.g. $\hat{p}(x, y)$ and $\hat{p}(x)$ for MI) are estimated with a *kernel function* Θ, which measures "similarity" between pairs of samples $\{x_n, y_n\}$ and $\{x_{n'}, y_{n'}\}$ using a resolution or *kernel width* r. For example, we can estimate:

$$\hat{p}_r(x_n, y_n) = \frac{1}{N} \sum_{n'=1}^{N} \Theta\left(\left| \begin{pmatrix} x_n - x_{n'} \\ y_n - y_{n'} \end{pmatrix} \right| - r \right). \tag{S.63}$$

By default Θ is the step kernel ($\Theta(x > 0) = 0$, $\Theta(x \leq 0) = 1$), and the norm $|\cdot|$ is the maximum distance. This combination – a *box kernel* – is what is implemented in JIDT. It results in $\hat{p}_r(x_n, y_n)$ being the proportion of the N values which fall within r of $\{x_n, y_n\}$ in both dimensions X and Y. Different resolutions r may be used for the different variables, whilst if using the same r then prior normalisation of the variables is sensible. Other choices for the kernel Θ and the norm $|\cdot|$ are possible. Conditional probabilities may be defined in terms of their component joint probabilities. These plug-in estimates for the PDFs are then used directly in evaluating a local measure for each sample $n \in [1, N]$ and averaging these over all samples, i.e. via Eq. (S.36) for $H(X)$ rather than via Eq. (S.1) (e.g. see **Kaiser and Schreiber** (2002) for transfer entropy). Note that methods for bias-correction here are available for

[11] This method can produce a local or pointwise Granger causality, as a local transfer entropy using a Gaussian model estimator.

individual entropy estimates (e.g. as proposed by **Grassberger** (1988) for the box kernel), but when combined for sums of entropies (as in MI, TE, etc.) **Kaiser and Schreiber** (2002) state: "this approach is not viable ... since the finite sample fluctuations ... are not independent and we cannot correct their bias separately".[12] Such issues are addressed by the Kraskov-Stögbauer-Grassberger estimator in the next section.

Kernel estimation can measure non-linear relationships and is model-free (unlike Gaussian estimators), though is sensitive to the parameter choice for resolution r (**Schreiber**, 2000; **Kaiser and Schreiber**, 2002) (see below), is biased and is less time-efficient. Naive algorithms require $\mathrm{O}\left(N^2\right)$ time, although efficient neighbour searching can reduce this to $\mathrm{O}\left(N \log N\right)$ or via box-assisted methods to $\mathrm{O}\left(N\right)$ (**Kantz and Schreiber**, 1997).[13] Box-assisted methods are used in JIDT for maximal efficiency.

Selecting a value for r can be difficult, with a too small value yielding undersampling effects (e.g. MI the values diverge (**Schreiber**, 2000)) whilst a too large value ignores subtleties in the data. One can heuristically determine a lower bound for r to avoid undersampling. Assuming all data are normalised (such that r then refers to a number of standard deviations) and spread somewhat evenly, the values for each variable roughly span 6 standard deviations and a given sample has $\sim N/(6/2r)$ coincident samples in any given dimension or $\sim N/(6/2r)^d$ in the full joint space of d dimensions. Requiring some number K of coincident samples on average within r (**Lungarella et al.** (2005) suggest $K \geq 3$ though at least 10 is more common), we then solve for $K \leq N/(6/2r)^d$.[14] Even within these extremes however, the choice of r can have a very large influence on the comparative results of the measure; see **Schreiber** (2000) and **Kaiser and Schreiber** (2002).

S.2.2.3 Kraskov-Stögbauer-Grassberger (KSG) technique **Kraskov, Stögbauer, and Grassberger** (2004) (KSG) (see also **Kraskov** (2004)) improved on (box) kernel estimation for MI by combining several specific enhancements designed to reduce errors when handling a small number of observations. These include: the use of Kozachenko-Leonenko estimators (**Kozachenko and Leonenko**, 1987) of log-probabilities via nearest-neighbour counting; bias correction; and a fixed number K of nearest-neighbours in the full X-Y joint space. The latter effectively means using a dynamically altered kernel width r to adjust to the density of samples in the vicinity of any given observation, which smooths out errors in the PDF estimation. For each sample $\{x, y\}$, one finds the Kth nearest neighbour in the full $\{x, y\}$ space (using max norms to compare x and y distances), and sets kernel widths r_x and r_y from it. The authors then propose two different algorithms for determining r_x and r_y from the Kth nearest neighbour.

For the first KSG algorithm, r_x and r_y are set to the maximum of the x and y distances to the Kth nearest neighbour, and one then counts the number of neighbours n_x and n_y strictly within these widths in each marginal space. Then the averages of n_x and n_y over all samples are used to compute:

$$I^{(1)}(X;Y) = \psi(K) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N), \tag{S.64}$$

(in *nats*) where ψ denotes the digamma function.

For the second KSG algorithm, r_x and r_y are set separately to the x and y distances to the Kth nearest neighbour, and one then counts the number of neighbours n_x and n_y within and on these widths in each

[12] As such, these are not implemented in JIDT, except for one method available for testing with the kernel estimator for TE.

[13] These quoted time complexities ignore the dependency on dimension d of the data, but will require a multiplier of at least d to determine norms, with larger multipliers perhaps required for more complicated box-assisted algorithms.

[14] More formally, one can consider the average number of coincidences for the *typical set*, see **Cover and Thomas** (1991) and **Marton and Shields** (1994).

marginal space. Again one uses the averages of n_x and n_y over all samples to compute (in *nats*):

$$I^{(2)}(X;Y) = \psi(K) - \frac{1}{K} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N). \tag{S.65}$$

Crucially, the estimator is bias corrected, and is demonstrated to be quite robust to variations in K (from $K = 4$ upwards, as variance in the estimate decreases with K) (**Kraskov et al.**, 2004). Of the two algorithms: algorithm 1 (Eq. (S.64)) is more accurate for smaller numbers of samples but is more biased, while algorithm 2 (Eq. (S.65)) is more accurate for very large sample sizes.

The KSG estimator is directly extendible to multi-information also; see **Kraskov** (2004).

Furthermore, **Kraskov** (2004) originally proposed that TE could be computed as the difference between two MIs (with each estimated using the aforementioned technique). However, the KSG estimation technique has since been properly extended to conditional MI by **Frenzel and Pompe** (2007) and transfer entropy (originally by **Gomez-Herrero et al.** (2010) and later for algorithm 2 by **Wibral et al.** (2014a)) with single estimators. Here for $I(X;Y \mid Z)$, for each sample $\{x, y, z\}$, one finds the Kth nearest neighbour in the full $\{x, y, z\}$ space (using max norms to compare x, y and z distances), and sets kernel widths r_x, r_y and r_z from it. Following KSG algorithm 1, r_z and $\{r_{xz}, r_{yz}\}$ are set to the maximum of the marginal distances to the Kth nearest neighbour, and one then counts $\{n_z, n_{xz}, n_{yz}\}$ strictly within this width (where n_{xz} and n_{yz} refer to counts in the joint $\{x, z\}$ and $\{y, z\}$ joint spaces) to obtain (**Frenzel and Pompe**, 2007; **Gomez-Herrero et al.**, 2010):

$$I^{(1)}(X;Y \mid Z) = \psi(K) + \langle \psi(n_z + 1) - \psi(n_{xz} + 1) - \psi(n_{yz} + 1) \rangle . \tag{S.66}$$

While following KSG algorithm 2, $\{r_x, r_y, r_z\}$ are set separately to the marginal distances to the Kth nearest neighbour, and one then counts $\{n_z, n_{xz}, n_{yz}\}$ within or on these widths to obtain (**Wibral et al.**, 2014a):

$$I^{(2)}(X;Y \mid Z) = \psi(K) - \frac{2}{K} + \left\langle \psi(n_z) - \psi(n_{xz}) + \frac{1}{n_{xz}} - \psi(n_{yz}) + \frac{1}{n_{yz}} \right\rangle . \tag{S.67}$$

Local values for these estimators can be extracted by unrolling the expectation values and computing the nearest neighbour counts only at the given observation $\{x, y\}$, e.g. for KSG algorithm 1 (**Lizier**, 2014):

$$i^{(1)}(x;y) = \psi(K) - \psi(n_x + 1) - \psi(n_y + 1) + \psi(N), \tag{S.68}$$

$$i^{(1)}(x;y \mid z) = \psi(K) + \psi(n_z + 1) - \psi(n_{xz} + 1) - \psi(n_{yz} + 1). \tag{S.69}$$

This approach has been used to estimate local transfer entropy by **Lizier et al.** (2011a) and **Steeg and Galstyan** (2013).

KSG estimation builds on the non-linear and model-free capabilities of kernel estimation to add bias correction, better data efficiency and accuracy, and being effectively parameter-free (being relatively stable to choice of K). As such, it is widely-used as best of breed solution for MI, conditional MI and TE for continuous data; see e.g. **Wibral et al.** (2014a) and **Vicente and Wibral** (2014). On the downside, it can be computationally expensive with naive algorithms requiring $\mathrm{O}\left(KN^2\right)$ time (again ignoring the dimensionality of the data) though fast nearest neighbour search techniques can reduce this to $\mathrm{O}\left(KN \log N\right)$. For release v1.0 JIDT only implements a naive algorithm, though fast nearest neighbour search is implemented and available via the project SVN, and as such will be included in future releases.

S.2.2.4 Permutation entropy and symbolic TE *Permutation entropy* approaches (**Bandt and Pompe**, 2002) estimate the relevant PDFs based on the relative ordinal structure of the joint vectors (this is not suitable for PDFs of single dimensional variables). That is, for a joint variable $\mathbf{X}$ of d dimensions, a sample $\mathbf{x}$ with components x_i ($i \in \{0 \ldots d-1\}$) is replaced by an ordinal vector $\mathbf{o}$ with components $o_i \in \{0 \ldots d-1\}$, where the value of $o_i = r$ assigned for x_i being the r-th largest component in $\mathbf{x}$. The PDF $p(\mathbf{x})$ is replaced by computation of $\hat{p}(\mathbf{o})$ for the corresponding ordinal vector, and these are used as plug-in estimates for the relevant expected or local information-theoretic measure.

Permutation entropy has for example been adapted to estimate TE as the *symbolic transfer entropy* (**Staniek and Lehnertz**, 2008), with local symbolic transfer entropy also defined (**Nakajima et al.**, 2012; **Nakajima and Haruna**, 2013).

Permutation approaches are computationally fast, since they effectively compute a discrete entropy after the ordinal symbolisation ($\mathrm{O}(N)$). They are a model-based approach however, assuming that all relevant information is in the ordinal relationship between the variables. This is not necessarily the case, and can lead to misleading results, as demonstrated by **Wibral et al.** (2013).

REFERENCES

Ash, R. B. (1965), Information Theory (Dover Publications Inc., New York)

Ay, N. and Polani, D. (2008), Information Flows in Causal Networks, *Advances in Complex Systems*, 11, 1, 17–41

Bandt, C. and Pompe, B. (2002), Permutation Entropy: A Natural Complexity Measure for Time Series, *Physical Review Letters*, 88, 17, 174102+, doi:10.1103/physrevlett.88.174102

Barnett, L. (2013), Personal Communication

Barnett, L., Barrett, A. B., and Seth, A. K. (2009a), Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables, *Physical Review Letters*, 103, 23, 238701+, doi:10.1103/physrevlett.103.238701

Barnett, L. and Bossomaier, T. (2012), Transfer Entropy as a Log-Likelihood Ratio, *Physical Review Letters*, 109, 138105+, doi:10.1103/physrevlett.109.138105

Barnett, L., Buckley, C. L., and Bullock, S. (2009b), Neural complexity and structural connectivity, *Physical Review E*, 79, 5, 051914+, doi:10.1103/physreve.79.051914

Bertschinger, N., Rauh, J., Olbrich, E., and Jost, J. (2013), Shared InformationNew insights and problems in decomposing information in complex systems, in T. Gilbert, M. Kirkilionis, and G. Nicolis, eds., Proceedings of the European Conference on Complex Systems 2012 (Springer, Switzerland), Springer Proceedings in Complexity, 251–269, doi:10.1007/978-3-319-00395-5_35

Bialek, W., Nemenman, I., and Tishby, N. (2001), Complexity through nonextensivity, *Physica A: Statistical Mechanics and its Applications*, 302, 1-4, 89–99, doi:10.1016/s0378-4371(01)00444-7

Boedecker, J., Obst, O., Lizier, J. T., Mayer, and Asada, M. (2012), Information processing in echo state networks at the edge of chaos, *Theory in Biosciences*, 131, 3, 205–213, doi:10.1007/s12064-011-0146-8

Bonachela, J. A., Hinrichsen, H., and Muñoz, M. A. (2008), Entropy estimates of small data sets, *Journal of Physics A: Mathematical and Theoretical*, 41, 20, 202001+, doi:10.1088/1751-8113/41/20/202001

Brillinger, D. R. (2004), Some data analyses using mutual information, *Brazilian Journal of Probability and Statistics*, 18, 163–183

Chávez, M., Martinerie, J., and Le Van Quyen, M. (2003), Statistical assessment of nonlinear causality: application to epileptic EEG signals, *Journal of Neuroscience Methods*, 124, 2, 113–128

Cheng, P. E., Liou, J. W., Liou, M., and Aston, J. A. D. (2006), Data information in contingency tables: A fallacy of hierarchical loglinear models, *Journal of Data Science*, 4, 4, 387–398

Chicharro, D. and Ledberg, A. (2012), When Two Become One: The Limits of Causality Analysis of Brain Dynamics, *PLoS ONE*, 7, 3, e32466+, doi:10.1371/journal.pone.0032466

Cover, T. M. and Thomas, J. A. (1991), Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing) (Wiley-Interscience, New York), 99 edition

Crutchfield, J. P. and Feldman, D. P. (2003), Regularities Unseen, Randomness Observed: Levels of Entropy Convergence, *Chaos*, 13, 1, 25–54, doi:10.1063/1.1530990

Crutchfield, J. P. and Young, K. (1989), Inferring statistical complexity, *Physical Review Letters*, 63, 2, 105–108, doi:10.1103/physrevlett.63.105

Dasan, J., Ramamohan, T. R., Singh, A., and Nott, P. R. (2002), Stress fluctuations in sheared Stokesian suspensions, *Physical Review E*, 66, 2, 021409

DeWeese, M. R. and Meister, M. (1999), How to measure the information gained from one symbol, *Network: Computation in Neural Systems*, 10, 325–340

Faes, L., Nollo, G., and Porta, A. (2011), Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique, *Physical Review E*, 83, 051112+, doi:10.1103/physreve.83.051112

Faes, L., Nollo, G., and Porta, A. (2012), Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series, *Computers in Biology and Medicine*, 42, 3, 290–297, doi:10.1016/j.compbiomed.2011.02.007

Fano, R. M. (1961), Transmission of information: a statistical theory of communications (M.I.T. Press, Cambridge, MA, USA)

Frenzel, S. and Pompe, B. (2007), Partial Mutual Information for Coupling Analysis of Multivariate Time Series, *Physical Review Letters*, 99, 20, 204101+, doi:10.1103/physrevlett.99.204101

Geweke, J. (1982), Measurement of linear dependence and feedback between multiple time series, *Journal of the American Statistical Association*, 77, 378, 304–313, doi:10.1080/01621459.1982.10477803

Gomez-Herrero, G., Wu, W., Rutanen, K., Soriano, M. C., Pipa, G., and Vicente, R. (2010), Assessing coupling dynamics from an ensemble of time series, arXiv:1008.0539

Grassberger, P. (1986), Toward a quantitative theory of self-generated complexity, *International Journal of Theoretical Physics*, 25, 9, 907–938

Grassberger, P. (1988), Finite sample corrections to entropy and dimension estimates, *Physics Letters A*, 128, 6-7, 369–373, doi:10.1016/0375-9601(88)90193-4

Griffith, V. and Koch, C. (2014), Quantifying Synergistic Mutual Information, in M. Prokopenko, ed., Guided Self-Organization: Inception, volume 9 of *Emergence, Complexity and Computation* (Springer, Berlin/Heidelberg), 159–190, doi:10.1007/978-3-642-53734-9_6

Harder, M., Salge, C., and Polani, D. (2013), Bivariate measure of redundant information, *Physical Review E*, 87, 012130+, doi:10.1103/physreve.87.012130

Helvik, T., Lindgren, K., and Nordahl, M. G. (2004), Local information in one-dimensional cellular automata, in P. M. A. Sloot, B. Chopard, and A. G. Hoekstra, eds., Proceedings of the International Conference on Cellular Automata for Research and Industry, Amsterdam, volume 3305 of *Lecture Notes in Computer Science* (Springer, Berlin/Heidelberg), volume 3305 of *Lecture Notes in Computer Science*, 121–130, doi:10.1007/978-3-540-30479-1_13

Ito, S., Hansen, M. E., Heiland, R., Lumsdaine, A., Litke, A. M., and Beggs, J. M. (2011), Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model, *PLoS ONE*, 6, 11, e27431+, doi:10.1371/journal.pone.0027431

Kaiser, A. and Schreiber, T. (2002), Information transfer in continuous processes, *Physica D*, 166, 1-2, 43–62

Kantz, H. and Schreiber, T. (1997), Nonlinear Time Series Analysis (Cambridge University Press, Cambridge, MA)

Kozachenko, L. and Leonenko, N. (1987), A statistical estimate for the entropy of a random vector, *Problems of Information Transmission*, 23, 9–16

Kraskov, A. (2004), 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)

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), Estimating mutual information, *Physical Review E*, 69, 6, 066138+, doi:10.1103/physreve.69.066138

Lindner, M., Vicente, R., Priesemann, V., and Wibral, M. (2011), TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy, *BMC Neuroscience*, 12, 1, 119+, doi:10.1186/1471-2202-12-119

Lizier, J., Heinzle, J., Soon, C., Haynes, J. D., and Prokopenko, M. (2011a), Spatiotemporal information transfer pattern differences in motor selection, *BMC Neuroscience*, 12, Suppl 1, P261+, doi:10.1186/1471-2202-12-s1-p261

Lizier, J. T. (2013), The Local Information Dynamics of Distributed Computation in Complex Systems, Springer Theses (Springer, Berlin / Heidelberg), doi:10.1007/978-3-642-32952-4

Lizier, J. T. (2014), Measuring the dynamics of information processing on a local scale in time and space, in M. Wibral, R. Vicente, and J. T. Lizier, eds., Directed Information Measures in Neuroscience (Springer, Berlin/Heidelberg), Understanding Complex Systems, 161–193, doi:10.1007/978-3-642-54474-3_7

Lizier, J. T., Flecker, B., and Williams, P. L. (2013), Towards a synergy-based approach to measuring information modification, in Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE) (IEEE), 43–51, doi:10.1109/alife.2013.6602430

Lizier, J. T., Heinzle, J., Horstmann, A., Haynes, J.-D., and Prokopenko, M. (2011b), Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity, *Journal of Computational Neuroscience*, 30, 1, 85–107, doi:10.1007/s10827-010-0271-2

Lizier, J. T., Pritam, S., and Prokopenko, M. (2011c), Information dynamics in small-world Boolean networks, *Artificial Life*, 17, 4, 293–314, doi:10.1162/artl_a_00040

Lizier, J. T. and Prokopenko, M. (2010), Differentiating information transfer and causal effect, *European Physical Journal B*, 73, 4, 605–615, doi:10.1140/epjb/e2010-00034-5

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2007), Detecting Non-trivial Computation in Complex Dynamics, in Almeida, L. M. Rocha, E. Costa, I. Harvey, and A. Coutinho, eds., Proceedings of the 9th European Conference on Artificial Life (ECAL 2007), volume 4648 of *Lecture Notes in Computer Science* (Springer, Berlin / Heidelberg), volume 4648 of *Lecture Notes in Computer Science*, 895–904, doi:10.1007/978-3-540-74913-4_90

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008), Local information transfer as a spatiotemporal filter for complex systems, *Physical Review E*, 77, 2, 026110+, doi:10.1103/physreve.77.026110

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2010), Information modification and particle collisions in distributed computation, *Chaos*, 20, 3, 037109+, doi:10.1063/1.3486801

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2012), Local measures of information storage in complex distributed computation, *Information Sciences*, 208, 39–54, doi:10.1016/j.ins.2012.04.016

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2014), A Framework for the Local Information Dynamics of Distributed Computation in Complex Systems, in M. Prokopenko, ed., Guided Self-Organization: Inception, volume 9 of *Emergence, Complexity and Computation* (Springer Berlin Heidelberg), 115–158, doi:10.1007/978-3-642-53734-9_5

Lungarella, M., Pegors, T., Bulwinkle, D., and Sporns, O. (2005), Methods for quantifying the informational structure of sensory and motor data, *Neuroinformatics*, 3, 3, 243–262

Lungarella, M. and Sporns, O. (2006), Mapping Information Flow in Sensorimotor Networks, *PLoS Computational Biology*, 2, 10, e144+, doi:10.1371/journal.pcbi.0020144

MacKay, D. J. C. (2003), Information Theory, Inference, and Learning Algorithms (Cambridge University Press, Cambridge)

Manning, C. D. and Schütze, H. (1999), Foundations of Statistical Natural Language Processing (The MIT Press, Cambridge, MA, USA)

Marton, K. and Shields, P. C. (1994), Entropy and the Consistent Estimation of Joint Distributions, *The Annals of Probability*, 22, 2, 960–977

Nakajima, K. and Haruna, T. (2013), Symbolic local information transfer, *The European Physical Journal Special Topics*, 222, 2, 437–455, doi:10.1140/epjst\%252fe2013-01851-x

Nakajima, K., Li, T., Kang, R., Guglielmino, E., Caldwell, D. G., and Pfeifer, R. (2012), Local information transfer in soft robotic arm, in 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO) (IEEE), 1273–1280, doi:10.1109/robio.2012.6491145

Obst, O., Boedecker, J., and Asada, M. (2010), Improving Recurrent Neural Network Performance Using Transfer Entropy Neural Information Processing. Models and Applications, in K. Wong, B. Mendis, and A. Bouzerdoum, eds., Neural Information Processing. Models and Applications, volume 6444 of *Lecture Notes in Computer Science* (Springer, Berlin/Heidelberg), chapter 24, 193–200, doi:10.1007/978-3-642-17534-3_24

Paninski, L. (2003), Estimation of entropy and mutual information, *Neural Computation*, 15, 6, 1191–1253, doi:10.1162/089976603321780272

Prokopenko, M. and Lizier, J. T. (2014), Transfer entropy and transient limits of computation, *Scientific Reports*, 4, 5394+, doi:10.1038/srep05394

Prokopenko, M., Lizier, J. T., and Price, D. C. (2013), On Thermodynamic Interpretation of Transfer Entropy, *Entropy*, 15, 2, 524–543, doi:10.3390/e15020524

Schreiber, T. (2000), Measuring Information Transfer, *Physical Review Letters*, 85, 2, 461–464, doi:10.1103/physrevlett.85.461

Shalizi, C. R. (2001), Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata, Ph.D. thesis, University of Wisconsin-Madison

Shalizi, C. R., Haslinger, R., Rouquier, J.-B., Klinkner, K. L., and Moore, C. (2006), Automatic filters for the detection of coherent structure in spatiotemporal systems, *Physical Review E*, 73, 3, 036104

Shannon, C. E. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27, 379–423 & 623–656

Staniek, M. and Lehnertz, K. (2008), Symbolic Transfer Entropy, *Physical Review Letters*, 100, 15, 158101+, doi:10.1103/physrevlett.100.158101

Steeg, G. V. and Galstyan, A. (2013), Information-theoretic measures of influence based on content dynamics, in Proceedings of the Sixth ACM international conference on Web search and data mining (ACM, New York, NY, USA), WSDM '13, 3–12, doi:10.1145/2433396.2433400

Stramaglia, S., Wu, G.-R., Pellicoro, M., and Marinazzo, D. (2012), 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 (IEEE), 3668–3671, doi:10.1109/embc.2012.6346762

Takens, F. (1981), Detecting strange attractors in turbulence, in D. Rand and L.-S. Young, eds., Dynamical Systems and Turbulence, Warwick 1980, volume 898 of *Lecture Notes in Mathematics* (Springer, Berlin / Heidelberg), chapter 21, 366–381, doi:10.1007/bfb0091924

Tononi, G., Sporns, O., and Edelman, G. M. (1994), 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, doi:10.1073/pnas.91.11.5033

Vakorin, V. A., Krakovska, O. A., and McIntosh, A. R. (2009), Confounding effects of indirect connections on causality estimation, *Journal of Neuroscience Methods*, 184, 1, 152–160

Verdes, P. F. (2005), Assessing causality from multivariate time series, *Physical Review E*, 72, 2, 026222+, doi:10.1103/physreve.72.026222

Vicente, R. and Wibral, M. (2014), Efficient estimation of information transfer, in M. Wibral, R. Vicente, and J. T. Lizier, eds., Directed Information Measures in Neuroscience (Springer, Berlin/Heidelberg), Understanding Complex Systems, 37–58, doi:10.1007/978-3-642-54474-3_2

Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011), Transfer entropya model-free measure of effective connectivity for the neurosciences, *Journal of Computational Neuroscience*, 30, 1, 45–67, doi:10.1007/s10827-010-0262-3

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., et al. (2013), Measuring Information-Transfer Delays, *PLoS ONE*, 8, 2, e55809+, doi:10.1371/journal.pone.0055809

Wibral, M., Vicente, R., and Lindner, M. (2014a), Transfer entropy in neuroscience, in M. Wibral, R. Vicente, and J. T. Lizier, eds., Directed Information Measures in Neuroscience (Springer, Berlin/Heidelberg), Understanding Complex Systems, 3–36, doi:10.1007/978-3-642-54474-3_1

Wibral, M., Vicente, R., and Lizier, J. T., eds. (2014b), Directed Information Measures in Neuroscience (Springer, Berlin, Heidelberg), doi:10.1007/978-3-642-54474-3

Williams, P. L. and Beer, R. D. (2010), Nonnegative Decomposition of Multivariate Information, arXiv:1004.2515

Williams, P. L. and Beer, R. D. (2011), Generalized Measures of Information Transfer, arXiv:1102.1507

Wollstadt, P., Martínez-Zarzuela, M., Vicente, R., Díaz-Pernas, F. J., and Wibral, M. (2014), Efficient transfer entropy analysis of Non-Stationary neural time series, *PLoS ONE*, 9, 7, e102833+, doi:10.1371/journal.pone.0102833