J Comput Neurosci (2014) 36:119–140
DOI 10.1007/s10827-013-0458-4

Synergy, redundancy, and multivariate information measures: an experimentalist's perspective

Nicholas Timme · Wesley Alford · Benjamin Flecker · John M. Beggs

Received: 9 August 2012 / Revised: 26 April 2013 / Accepted: 29 April 2013 / Published online: 3 July 2013

Abstract Information theory has long been used to quantify interactions between two variables. With the rise of complex systems research, multivariate information measures have been increasingly used to investigate interactions between groups of three or more variables, often with an emphasis on so called synergistic and redundant interactions. While bivariate information measures are commonly agreed upon, the multivariate information measures in use today have been developed by many different groups, and differ in subtle, yet significant ways. Here, we will review these multivariate information measures with special emphasis paid to their relationship to synergy and redundancy, as well as examine the differences between these measures by applying them to several simple model systems. In addition to these systems, we will illustrate the usefulness of the information measures by analyzing neural spiking data from a dissociated culture through early stages of its development. Our aim is that this work will aid other researchers as they seek the best multivariate information measure for their specific research goals and system. Finally, we have made software available online which allows the user to calculate all of the information measures discussed within this paper.

Keywords Information theory · Multivariate information measures · Complex systems · Neural coding · Dissociated neuronal cultures · Multielectrode array

Action Editor: Jonathan David Victor

N. Timme (✉) · W. Alford · B. Flecker · J. M. Beggs
Department of Physics, Indiana University,
Bloomington, IN 47405-7105, USA
e-mail: nmtimme@umail.iu.edu

1 Introduction

Since its introduction by Shannon to quantify communication (Shannon 1948), information theory has proved to be a useful tool in many disciplines. It has been successfully applied in several areas of research, including neuroscience (Rieke et al. 1997), data compression (Ziv and Lempel 1977), coding (Berrou et al. 1993), dynamical systems (Fraser and Swinney 1986), and genetic coding (Butte and Kohane 2000), just to name a few. Information theory's broad applicability is due in part to the fact that it relies only on the probability distribution associated with one or more variables. Generally speaking, information theory uses the probability distributions associated with the values of the variables to ascertain whether or not the values of the variables are related and, depending on the situation, the way in which they are related. As a result of this, information theory can be applied to linear and non-linear systems. In summary, information theory is a model-independent approach.

Information theoretic approaches to problems involving one and two variables are well understood and widely used. However, many systems contain interactions between three or more variables (in neuroscience see, for instance, Quiroga and Panzeri 2009; Ohiorhenuan 2010; Yeh et al 2010). Several information measures have been introduced to analyze these multivariate interactions (McGill 1954; Watanabe 1960; Han 1975; Chechik et al. 2001; Nirenberg et al. 2001; Schneidman et al. 2003b; Varadan et al. 2006; Williams and Beer 2010). Frequently, these measures were introduced to measure so called "synergy" and/or "redundancy." These multivariate information measures have been applied in physical systems (Cerf and Adami 1997; Matsuda 2000), biological systems (Anastassiou 2007; Chanda et al. 2007), and neuroscience (Brenner et al. 2000;

Schneidman et al. 2003a; Bettencourt et al. 2007). However, these multivariate information measures differ in significant and sometimes subtle ways. Furthermore, the notation and naming associated with these measures is inconsistent throughout the literature (see for example, Watanabe 1960; Tanoni et al. 1994; Sporns et al. 2000; Schneidman et al. 2003b; Wennekers and Ay 2003).

Within this paper, we will examine a wide array of multivariate information measures in an attempt to clearly articulate the different measures and their uses. First, we will discuss the concepts of synergy and redundancy, as well as how these multivariate information measures fit into data analysis. Then, we will review the information theory behind each individual measure while paying specific attention to two perspectives of interactions: those that exist within a group of variables and those that exist between a group of variables and another target variable. Next, we will apply the information measures to several model systems in order to illuminate their differences and similarities. Also, we will apply the information measures to neural spiking data from a dissociated neural culture. Finally, we will discuss the overall results for each information measure in turn, as well as how the information measures could be used in three example experiments in neuroscience. Our goal is to clarify these methods for other researchers as they search for the multivariate information measure that will best address their specific research goals. We wish to emphasize that we are not attempting to guide the development of information theory; rather we seek to facilitate a broader use of information theoretic tools.

In order to facilitate the use of the information measures discussed in this paper, we have made freely available MATLAB software that can be used to calculate all of the information measures discussed herein.[1] An earlier version of this work was previously posted on the arXiv (Timme et al. 2011).

2 Information theory analysis

Before proceeding to a discussion of the information measures, we will briefly note the place of these measures in the various types of data analyses that can be performed, as well as the types of experimental questions that can be addressed with these analyses (see Victor 2006; Quiroga and Panzeri 2009 for further reviews in neuroscience).

Information theory can be applied to many types of systems to address many types of experimental questions. Often times, information theory is employed to examine whether or not and to what degree one or more variables encode information about some other variable(s). In neuroscience, information theory is typically applied in two widely-used experimental systems: *in vivo* sensory encoding experiments (see Optican and Richmond 1987; Borst and Theunissen 1999; Brenner et al. 2000; Panzeri et al. 2001; Rokem et al. 2006; Gollisch and Meister 2008 as examples) and network experiments (see Butts and Rokhsar 2001; Bettencourt et al. 2007, 2008; Garofalo et al. 2009 as examples).

Sensory encoding experiments typically involve the simultaneous recording of a controlled stimulus and the resulting neural activity of a subject. Experimental questions in this area might include, but are not limited to: Do individual neurons encode certain features of a stimulus or do neurons somehow work together to encode the stimulus? What feature of the neural activity contains the code (e.g. spike rate or spike timing)? How robust is the encoding process to different types of noise? How do the measurements resulting from information theory relate to the physical structure of the sensory processing system in the organism? (See Quiroga and Panzeri 2013 for a recent review of the issues related coding.)

We would like to emphasize that encoding experiments are fundamentally different from decoding experiments. Sensory decoding experiments involve the construction of a function that takes the neural activity as its input and produces an estimate of the stimulus variable as an output (Bialek et al. 1991; Warland et al. 1997; Pillow et al. 2008). There may be many ways (e.g. linear filter or generalized linear/nonlinear models) to construct this function. Information theory does not specify how this should be done and does not specify how to measure the error between the estimated and actual stimulus variables (Schneidman et al. 2003a). Because decoding is often done without reference to information theory, we will focus on encoding and network experiments as examples.

Network experiments typically involve the simultaneous recording of several neural sources. Experimental questions in this area might include, but are not limited to: Can the activity of a neural source be used to predict the activity of another neural source, thus implying a functional or effective connection between the two sources? How do the graphical models of the neural sources that can be generated relate to the effective, functional, or physical connectivity of the entire system? What types of interactions are present in the network and to what degree? How much of the neural activity can be accounted for by different subsets of sources?

Notice that, since information theory utilizes probability distributions, encoding experiments involving n variables and network experiments involving the spontaneous activity of n variables are equivalent in terms of the structure of the probability distribution and thus indistinguishable in terms of the application of information theory. For instance, a researcher may record action potentials from two

[1] http://mypage.iu.edu/~nmtimme.

neurons in the visual system of an animal as well as some visual stimulus variable. Another researcher may record action potentials from three neurons using a multi-electrode array. Both researchers would be interested in understanding the interactions between the three variables in his or her respective experiment. Either researcher can apply any of the multivariate information measures discussed herein to the probability distributions from their systems. Clearly, the choice of the information measure and the way in which it is applied to the system will directly impact the conclusions that can be drawn from the analysis. However, in terms of the structure of the necessary probability distributions, these two types of experiments are equivalent. As a result, they can be treated equally with the multivariate information measures discussed herein.

In the typical analysis, raw data is gathered from some source, such as neurons, an experimentally controlled variable, stock values, genes, or some other source. Then, the data are processed (e.g. spike detection in single neuron recordings) and converted to joint probability distributions. Depending on the type of data being analyzed, this process often involves binning the data into discrete states and discrete temporal units. Also, depending on the causal relationship between the elements of the experimental system, a time delay may be introduced between the variables. Once these joint probability distributions are obtained, they are passed through the chosen information measure to obtain a final result.

As all experimentalists know, the choices made throughout an analysis can dramatically affect the final results and errors early in the process can propagate throughout the analysis with unpredictable effects. How do you choose which data to analyze? The data are often continuous and always noisy and limited, so how do you know if you are applying the right binning algorithm? Even if the data were perfect and infinite, how do you conceptually choose the correct binning algorithm? Furthermore, could a binless algorithm be utilized to avoid binning problems (Victor 2002; Paiva et al. 2010)? Finally, even if you obtain the "best" possible probability distribution, how do you choose the right information measure to answer your experimental question? Within this paper we will only seek to address this final question, though we wish to emphasize that a careful consideration of all these points is *essential* to any analysis.

The continuous, noisy, and limited nature of data can affect the results from information measures in complex ways. Several efforts have been made to take these effects into account theoretically for some, *but not all*, of the information measures discussed herein (Treves and Panzeri 1995; Panzeri and Treves 1996; Strong et al. 1997; Paninski 2003; Schneidman et al. 2003a; Nemenman et al. 2004; Kennel et al. 2005; Averbeck et al. 2006; Hlaváčková-Schindler et al. 2007; Panzeri et al. 2007; Shlens et al. 2007). Generally, these works discuss correction terms that can be applied to information values and/or approximation methods that can be applied to information values to correct for sampling bias. As far as we are aware, research on the bias associated with an information measure has only been conducted on the most basic information theory measures we consider herein: entropy and mutual information (see Panzeri and Treves 1996; Strong et al. 1997 in particular). Several of the information measures we will discuss can be directly defined in terms of entropy and mutual information, but several cannot. Given the complexity of this topic and the fact that the bias associated with several of the information measures discussed herein have not been studied, a thorough examination of the bias produced by the sampling of noisy data is beyond the scope of this review. Though, we will discuss the issue briefly in connection with an example data analysis in Section 5.7. In conclusion, the reader is cautioned to carefully consider the effects of sampling bias in his or her analysis.

3 Synergy and redundancy

A crucial topic related to multivariate information measures is the distinction between synergy and redundancy. Many of the proposed information measures purport to measure synergy or redundancy, though the precise meanings of "synergy" and "redundancy" have not been agreed upon (see, for instance, Brenner et al. 2000; Williams and Beer 2010). (For a recent treatment of synergy in this context, see Griffith and Koch 2012.)

To begin to understand synergy, we can use a simple system. Suppose two variables (call them X_1 and X_2) provide some information about a third variable (call it Y). In other words, if you know the state of X_1 and X_2, then you know something about the state Y. Loosely said, the portion of that information that is not provided by knowing both X_1 alone and X_2 alone is said to be provided synergistically by X_1 and X_2. The synergy is the bonus information received by knowing X_1 and X_2 *together*, instead of separately.

We can take a similar initial approach to redundancy. Again, suppose X_1 and X_2 provide some information about Y. The common portion of the information X_1 provides alone and the information X_2 provides alone is said to be provided redundantly by X_1 and X_2. The redundancy is the information received from *both* X_1 and X_2.

These imprecise definitions may seem clear enough, but in attempting to measure these quantities, researchers have created distinct measures that produce different results. Based on the fact that the overall goal has not been clearly defined, it cannot be said that one of these measures is "correct." Rather, each measure has its own uses and limitations. Using the simple systems below, we will

attempt to clearly articulate the differences between the multivariate information measures.

4 Multivariate information measures

In this section we will discuss the various multivariate information theoretic measures that have been introduced previously. Of special note is the fact that the names and notation used in the literature have not been consistent. We will attempt to clarify the discussion as much as possible by listing alternative names when appropriate. We will refer to an information measure by its original name (or at least, its original name to the best of our knowledge).

4.1 Entropy and mutual information

The information theoretic quantities involving one and two variables are well-defined and their results are well-understood. Regarding the probability distribution of one variable (call it $p(x)$), the canonical measure is the entropy $H(x)$ (Cover and Thomas 2006). The entropy is given by: [2]

$$H(X) \equiv -\sum_{x \in X} p(x) \log(p(x)) \tag{1}$$

The entropy quantifies the amount of uncertainty that is present in the probability distribution. If the probability distribution is concentrated near one value, the entropy will be low. If the probability distribution is uniform, the entropy will be maximized.

When examining the relationship between two variables, the mutual information (I) quantifies the amount of information provided about one of the variables by knowing the value of the other (Cover and Thomas 2006). The mutual information is given by:

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{2}$$

where the conditional entropy is given by:

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} p(y) H(X|y) \\ &= \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log \frac{1}{p(x|y)} \end{aligned} \tag{3}$$

[2] Throughout the paper we will use capital letters to refer to variables and lower case letters to refer to individual values of those variables. We will also use discrete variables, though several of the information measures discussed can be directly extended to continuous variables. When working with a continuous variable, various techniques exists, such as kernel density estimation, which can be used to infer a discrete distribution from a continuous variable. Logarithms will be base 2 throughout in order to produce information values in units of bits.

The mutual information can also be written as the Kullback-Leibler divergence between the joint probability distribution of the actual data and the joint probability distribution of the independent model (wherein the joint distribution is equal to the product of the marginal distributions). This form is given by:

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{4}$$

The mutual information can be used as a measure of the interactions among more than two variables by grouping the variables into sets and treating each set as a single vector-valued variable. In this way, the mutual information can be used to measure the interactions between a group of variables and a target variable. For instance, the mutual information can be calculated between Y and the set $S = \{X_1, X_2\}$[3] in the following way:

$$I(Y;S) = \sum_{\substack{y \in Y \\ x_1 \in X_1, x_2 \in X_2}} p(y, x_1, x_2) \log \left(\frac{p(y, x_1, x_2)}{p(y)p(x_1, x_2)} \right) \tag{5}$$

However, when the mutual information is considered as in Eq. (5), it is not possible to separate contributions from individual X variables in the set S. Still, by varying the number of variables in S, the mutual information in Eq. (5) can be used to measure the gain or loss in information about Y by those variables in S. Along these lines, Bettencourt et al. used the mutual information between one variable (in their case, the activity of a neuron) and many other variables considered together (in their case, the activities of a group of other neurons) in order to examine the relationship between the amount of information the group of neurons provided about the single neuron to the number of neurons considered in the group (Bettencourt et al. 2008).

The mutual information can be conditioned upon a third variable to yield the conditional mutual information (Cover and Thomas 2006). It is given by:

$$\begin{aligned} &I(X;Y|Z) \\ &= \sum_{z \in Z} p(z) \sum_{x \in X, y \in Y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \sum_{x \in X, y \in Y, z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \end{aligned} \tag{6}$$

The conditional mutual information quantifies the amount of information one variable provides about a second variable

[3] We will use S to refer to a set of n X variables such that $S = \{X_1, X_2, \ldots X_n\}$ throughout the paper.

when a third variable is known. Note that it is possible for the conditional mutual information $I(X;Y|Z)$ to be larger or smaller than the mutual information $I(X;Y)$.

4.2 Interaction information

The first attempt to quantify the relationship among three variables in a joint probability distribution was the interaction information (II), which was introduced by McGill (McGill 1954). It attempts to extend the concept of the mutual information as the information gained about one variable by knowing the other. The interaction information is given by:

$$\begin{aligned} II(X;Y;Z) &\equiv I(X;Y|Z) - I(X;Y) \\ &= I(X;Z|Y) - I(X;Z) \\ &= I(Z;Y|X) - I(Z;Y) \end{aligned} \tag{7}$$

Given the fact that the conditional mutual information can be larger or smaller than the mutual information for the same set of variables, the interaction information can be positive or negative. Of the interaction information, McGill said (McGill 1954), "We see that $II(X;Y;Z)$ is the gain (or loss) in sample information transmitted between any two of the variables, due to the additional knowledge of the third variable." The interaction information can also be written as:

$$II(X;Y;Z) = I(X,Y;Z) - (I(X;Z) + I(Y;Z)) \tag{8}$$

In the form given in Eq. (8), the interaction information has been widely used in the literature and has often been referred to as the synergy (Gat and Tishby 1999; Brenner et al. 2000; Schneidman et al. 2003a; Anastassiou 2007) and the redundancy-synergy index (Chechik et al. 2001). Some authors have used the term "synergy" because they have interpreted a positive interaction information result to imply a synergistic interaction among the variables and a negative interaction information result to imply a redundant interaction among the variables. Thus, if we assume this interpretation of the interaction information and that the interaction information correctly measures multivariate interactions, then synergy and redundancy are taken to be *mutually exclusive* qualities of the interactions between variables. This view will find a counterpoint in the partial information decomposition to be discussed below.

Note that the interaction information does not theoretically differentiate between its three input variables. Equation (8) is structured in such a way that it appears that variables X and Y are being related to Z, but, based on Eq. (7), we see that we can permute the variables at will. Thus, the interaction information measures the interactions among a group of variables, as opposed to the interactions between a group of variables and a target variable.

The interaction information can also be written as an expansion of the entropies and joint entropies of the variables:

$$\begin{aligned} II(X;Y;Z) = &-H(X) - H(Y) - H(Z) + H(X,Y) \\ &+H(X,Z) + H(Y,Z) - H(X,Y,Z) \end{aligned} \tag{9}$$

This form leads to an expansion for the interaction information for n number of variables (Jakulin and Bratko 2008). If $S = \{X_1, X_2, \ldots X_n\}$, then the interaction information becomes:

$$II(S) = -\sum_{T \subseteq S} (-1)^{|S|-|T|} H(T) \tag{10}$$

In Eq. (10), T is a subset of S and $|S|$ denotes the set size of S. From Eq. (10), it is apparent that the interaction information treats all input variables equally and measures interactions among all variables for any number of input variables.

A measure similar to the interaction information was introduced by Bell and is referred to as the co-information (CI) (Bell 2003). It is given by the following expansion:

$$CI(S) \equiv -\sum_{T \subseteq S} (-1)^{|T|} H(T) = (-1)^{|S|} II(S) \tag{11}$$

Clearly, the co-information is equal to the interaction information when S contains an even number of variables and is equal to the negative of the interaction information when S contains an odd number of variables. So, for the three variable case, the co-information becomes:

$$\begin{aligned} CI(X;Y;Z) &= I(X;Y) - I(X,Y|Z) \\ &= I(X;Z) + I(Y;Z) - I(X,Y;Z) \end{aligned} \tag{12}$$

Because the co-information is directly related to the interaction information for systems with any number of variables, we will forgo presenting results from the co-information. The co-information has also been referred to as the generalized mutual information (Matsuda 2000).

4.3 Total correlation

The interaction information finds its conceptual base in extending the idea of the mutual information as the information gained about a variable when the other variable is known. Alternatively, we could extend the idea of the mutual information as the Kullback-Leibler divergence between the joint distribution and the independent model. If we do this, we arrive at the total correlation (TC) introduced by Watanabe (Watanabe 1960). It is given by:

$$TC(S) \equiv \sum_{\vec{x} \in S} p(\vec{x}) \log \left(\frac{p(\vec{x})}{p(x_1)p(x_2)\ldots p(x_n)} \right) \tag{13}$$

In Eq. (13), $\vec{x}$ is a vector containing individual states of the X variables. As with the interaction information, the total correlation also treats all input variables equally, thus it measures interactions among a group of variables.

The total correlation can also be written in terms of entropies as:

$$TC(S) = \left(\sum_{X_i \in S} H(X_i) \right) - H(S) \tag{14}$$

In this form, the total correlation has been referred to as the multi-information (Schneidman et al. 2003b), the spatial stochastic interaction (Wennekers and Ay 2003), and the integration (Tononi et al. 1994; Sporns et al. 2000). Using Eq. (2), the total correlation can also be written using a series of mutual information terms (see Appendix A for more details):

$$\begin{aligned} TC(S) &= I(X_1; X_2) + I(X_1, X_2; X_3) \\ &\quad + \ldots I(X_1, \ldots, X_{n-1}; X_n) \end{aligned} \tag{15}$$

4.4 Dual total correlation

After the total correlation was introduced, a measure with a similar structure, called the dual total correlation (DTC), was introduced by Han (Han 1975; 1978). The dual total correlation is given by:

$$DTC(S) \equiv \left(\sum_{X_i \in S} H(S/X_i) \right) - (n-1)H(S) \tag{16}$$

In Eq. (16), S/X_i is the set S with X_i removed and n is the number of X variables in S. As with the total correlation, the dual total correlation also measures the interactions within a group of variables and treats all input variables equally.

The dual total correlation can also be written as (Abdallah and Plumbley 2010):

$$DTC(S) = H(S) - \sum_{X_i \in S} H(X_i|S/X_i) \tag{17}$$

The dual total correlation calculates the amount of entropy present in S beyond the sum of the entropies for each variable conditioned upon all other variables. The dual total correlation has also been referred to as the excess entropy (Olbrich et al. 2008) and the binding information (Abdallah and Plumbley 2010). Using Eqs. (2), (14), and (16), the dual total correlation can also be related to the total correlation by (see Appendix B for more details):

$$DTC(S) = \left(\sum_{X_i \in S} I(S/X_i; X_i) \right) - TC(S) \tag{18}$$

4.5 ΔI

A distinct information measure, called ΔI, was introduced by Nirenberg and Latham (Nirenberg et al. 2001; Latham and Nirenberg 2005). It was introduced to measure the importance of correlations in neural coding. For the purposes of this paper, we can apply ΔI to the following situation: consider some set of X variables (call this set S). The values of the variables in S are related in some way to the value of another variable (call it Y). In Nirenberg and Latham's original work, the X variables were signals from neurons and the Y variable was the value of some stimulus variable. ΔI compares the true probability distributions associated with these variables to one that assumes the X variables act independently (i.e., there are no correlations between the X variables beyond those that can be explained by Y). If these distributions are similar, then it can be assumed that there are no relevant correlations between the X variables. If, on the other hand, these distributions are not similar, then we can conclude that relevant correlations are present between the X variables.

The independent model assumes that the X variables act independently, so we can form the probability for the X states conditioned upon the Y variable state using a simple product:

$$p_{ind}(\vec{x}|y) = \prod_i p(x_i|y) \tag{19}$$

Then, the conditional probability of the Y variable on the X variables can be found using Bayes' theorem.

$$p_{ind}(y|\vec{x}) = \frac{p_{ind}(\vec{x}|y)p(y)}{p_{ind}(\vec{x})} \tag{20}$$

The independent joint distribution of the X variables is given by:

$$p_{ind}(\vec{x}) = \sum_{y \in Y} p_{ind}(\vec{x}|y)p(y) \tag{21}$$

Then, ΔI is given by the weighted Kullback-Leibler distance between the conditional probability of the Y variable on the X variables for the independent model and the actual conditional probability of the same type.

$$\Delta I(S; Y) \equiv \sum_{\vec{x} \in S} p(\vec{x}) \sum_{y \in Y} p(y|\vec{x}) \log \left(\frac{p(y|\vec{x})}{p_{ind}(y|\vec{x})} \right) \tag{22}$$

About ΔI, Nirenberg and Latham say (Latham and Nirenberg 2005), "[s]pecifically, ΔI is the cost in yes/no questions for not knowing about correlations: if one were guessing the value of the Y variable based on the X variables, $\vec{x}$, then it would take, on average, ΔI more questions to guess the value of Y if one knew nothing about the correlations than if one knew everything about them [Variable

names changed to match this work]." Unlike the interaction information, ΔI is restricted to be non-negative (Latham and Nirenberg 2005).

Note that, because it measures interactions between a group of variables and another variable, ΔI is fundamentally different from the multivariate information measures discussed so far. The previous measures treat all of the variables equally and are designed to measure interactions among all of the variables, whereas ΔI and the following information measures treat the X variables and the Y variable separately. Because of this, ΔI *and the following information measures are differently situated in their ability to assess the degree to and manner in which the X variables in S encode the Y variable*.

4.6 Redundancy-synergy index

Another multivariate information measure was introduced by Chechik et al. 2001. This measure was originally referred to as the redundancy-synergy index (RSI) and it was created as an extension of the interaction information. It is given by:

$$RSI(S;Y) \equiv I(S;Y) - \sum_{X_i \in S} I(X_i;Y) \tag{23}$$

The redundancy-synergy index is designed to be maximal and positive when the variables in S are purported to provide synergistic information about Y. It should be negative when the variables in S provide redundant information about Y. Notice that, like ΔI, the redundancy-synergy index measures the interactions between a group of variables and another variable, except when S contains two variables, in which case the redundancy-synergy index is equal to the interaction information. The redundancy-synergy index has been referred to as the SynSum (Globerson et al. 2009), the WholeMinusSum synergy (Griffith and Koch 2012), and the negative of the redundancy-synergy index has also been referred to as the redundancy (Schneidman et al. 2003b).

4.7 Varadan's synergy

Yet another multivariate information measure was introduced by Varadan et al. (2006). In the original work, this measure was referred to as the synergy, but to avoid confusing it with other measures, we will refer to this measure as Varadan's synergy (VS). It is given by:

$$VS(S;Y) \equiv I(S;Y) - max \sum_{j} I(S_j;Y) \tag{24}$$

In Eq. (24), S_j refers to the possible sub-sets of S and the maximum seeks the partition of S that produces the largest sum of mutual informations between the subsets of S and Y. So, for instance, if $S = \{X_1, X_2, X_3\}$, Varadan's synergy would be given by:

$$VS(S;Y) = I(S;Y) - max \begin{cases} I(X_1;Y) + I(X_2,X_3;Y) \\ I(X_2;Y) + I(X_1,X_3;Y) \\ I(X_3;Y) + I(X_1,X_2;Y) \\ I(X_1;Y) + I(X_2;Y) + I(X_3;Y) \end{cases} \tag{25}$$

Similar to the interaction information, when Varadan's synergy is positive, the variables in S are said to provide synergistic information about Y, while when Varadan's synergy is negative, the variables in S are said to provide redundant information about Y. Like the redundancy-synergy index, Varadan's synergy measures the interactions between a group of variables and another variable, except when S contains two variables, in which case Varadan's synergy is equal to the interaction information.

4.8 Partial information decomposition

Finally, we will examine the collection of information measures introduced by Williams and Beer in the partial information decomposition (PID) (Williams and Beer 2010). (See (James et al. 2011; Flecker et al. 2011; Griffith and Koch 2012; Lizier et al. 2013) for recent works involving the partial information decomposition). The partial information decomposition is a method of dissecting the mutual information between a set of variables S and one other variable Y into non-overlapping terms. These terms quantify the information provided by the set of variables in S about Y uniquely, redundantly, synergistically, and in mixed forms. Thus, the partial information decomposition measures the interactions between a single variable and a group of variables.

The partial information decomposition has several potential advantages over other measures. First, it produces only non-negative results, unlike the interaction information, the redundancy-synergy index, and Varadan's synergy. Second, it treats synergy and redundancy as conceptually different quantities that can be measured simultaneously in an interaction, while the interaction information and ΔI combine synergy and redundancy to each produce one value. Third, the partial information decomposition measures the interactions between the Y variable and individual members and subsets of S, unlike the other measures that treat the X variables together. However, unlike the other measures, the partial information decomposition produces many terms, the number of which diverges quickly with increased number of X variables, and these terms can be difficult to interpret.

For the sake of brevity, we will not describe the entire partial information decomposition here, but we will describe

the case where $S = \{X_1, X_2\}$. A description of the general case can be found in Williams and Beer's original work (Williams and Beer 2010). The relevant mutual informations are equal to sums of the partial information terms. For the case of two X variables, there are only four possible terms. Information about Y can be provided uniquely by each X variable, redundantly by both X variables, or synergistically by both X variables together. Written out, the relevant mutual informations are given by the following sums:

$$\begin{aligned} I(X_1, X_2; Y) \equiv\ & Synergy(Y; X_1, X_2) + Unique(Y; X_1) \\ & + Unique(Y; X_2) \\ & + Redundancy(Y; X_1, X_2) \end{aligned} \tag{26}$$

$$I(X_1; Y) \equiv Unique(Y; X_1) + Redundancy(Y; X_1, X_2) \tag{27}$$

$$I(X_2; Y) \equiv Unique(Y; X_2) + Redundancy(Y; X_1, X_2) \tag{28}$$

The relevant mutual information values can be calculated easily. As described by Williams and Beer, the redundancy term is equal to a new information expression: the minimum information function. This function attempts to capture the intuitive view that the redundant information for a given state of Y is the information that is contributed by both X variables about that state of Y (consult Williams and Beer's original work (Williams and Beer 2010) for details and further motivation). The minimum information function is related to the specific information (DeWeese and Meister 1999).[4] The specific information is given by:

$$I_{spec}(y; X) = \sum_{x \in X} p(x \mid y) \left[log\left(\frac{1}{p(y)}\right) - log\left(\frac{1}{p(y \mid x)}\right) \right] \tag{29}$$

In Eq. (29), the specific information quantifies the amount of information provided by X about a specific state of the Y variable.

The minimum information can then be calculated by comparing the amount of information provided by the different X variables for each state of the Y variable considered individually.

$$I_{min}(Y; X_1, X_2) \equiv \sum_{y \in Y} p(y) min_{X_i} I_{spec}(y; X_i) \tag{30}$$

The minimum in Eq. (30) is taken over each X variable considered separately. As described by Williams and Beer, the minimum information is equal to the redundancy:

$$Redundancy(Y; X_1, X_2) \equiv I_{min}(Y; X_1, X_2) \tag{31}$$

Note, there is a subtle but crucial difference between the minimum information function and the idea that the redundancy is the information contributed by both X variables. The minimum information is the average minimum *amount* of information about Y that can be obtained from any of the X variables. See Section 5.3 for an example of this distinction.

Once the redundancy term is calculated via the minimum information function, the remaining partial information terms can be calculated with ease for the two X variable case:

$$\begin{aligned} Synergy(Y; X_1, X_2) = & I(Y; X_1, X_2) - I(Y; X_1) - I(Y; X_2) \\ & + Redundancy(Y; X_1, X_2) \end{aligned} \tag{32}$$

$$Unique(Y; X_1) = I(Y; X_1) - Redundancy(Y; X_1, X_2) \tag{33}$$

$$Unique(Y; X_2) = I(Y; X_2) - Redundancy(Y; X_1, X_2) \tag{34}$$

It should also be noted that the partial information decomposition provides an explanation for negative interaction information values (Williams and Beer 2010). To see this, insert the partial information expansions in Eqs. (26), (27), and (28) into the mutual information terms in the interaction information (8):

$$\begin{aligned} II(X_1; X_2; Y) &= I(X_1, X_2; Y) - I(X_1, Y) - I(X_2, Y) \\ &= Synergy(Y; X_1, X_2) \\ &\quad - Redundancy(Y; X_1, X_2) \end{aligned} \tag{35}$$

Thus, the partial information decomposition finds that a negative interaction information value implies that the redundant contribution is greater in magnitude than the synergistic contribution. Furthermore, the structure of the partial information decomposition implies that synergistic and redundant interactions are not mutually exclusive, as was the case for the traditional interpretation of the interaction information. Thus, according to the partial information decomposition, there may be non-zero synergistic and redundant contributions simultaneously.

Throughout the remainder of this article, we will label the various terms in the partial information decomposition in accordance with the notation used by Williams and Beer. The term that has been interpreted as the synergy will be referred to as $\Pi_R(Y; \{12\})$ or PID synergy. The term that has been interpreted as the redundancy will be labeled as

[4] It should be noted that DeWeese and Meister refer to the expression in Eq. (29) as the specific surprise.

$\Pi_R(Y;\{1\}\{2\})$ or PID redundancy. The unique information terms will be referred to as $\Pi_R(Y;\{1\})$ and $\Pi_R(Y;\{2\})$, or simply as PID unique information.

When the partial information decomposition is extended to the case where $S = \{X_1, X_2, X_3\}$, new mixed terms are introduced to the expansions of the mutual informations. For instance, information can be supplied about Y redundantly between X_3 and the synergistic contribution from X_1 and X_2 (this term is noted as $\Pi_R(Y;\{12\}\{3\})$). In total, the partial information decomposition contains 18 terms when S contains three variables. To aide with the visualization of the partial information terms, Williams and Beer introduced partial information diagrams. The diagrams for the two and three X variable cases are shown in Figs. 1 and 2. The various synergy terms in the partial information decomposition have been referred to as S_{max} (Griffith and Koch 2012).

For the three X variable case, it can be shown that the interaction information between Y and the X variables contained in S is related to the partial information terms by the following equation (Williams and Beer 2010):

$$\begin{aligned} II(Y;X_1;X_2;X_3) = {} & \Pi_R(Y;\{123\}) + \Pi_R(Y;\{1\}\{2\}\{3\}) \\ & - \Pi_R(Y;\{1\}\{23\}) - \Pi_R(Y;\{2\}\{13\}) \\ & - \Pi_R(Y;\{3\}\{12\}) - \Pi_R(Y;\{12\}\{13\}) \\ & - \Pi_R(Y;\{12\}\{23\}) - \Pi_R(Y;\{13\}\{23\}) \\ & - 2\Pi_R(Y;\{12\}\{13\}\{23\}) \end{aligned} \tag{36}$$

From Eq. (36), we can see that the four-way interaction information is related to the partial information decomposition via a complicated summation of terms.

4.9 Additional methods

Beyond the measures discussed so far, additional methods have been proposed to measure the synergy and redundancy. Globerson et. al. introduced the minimum information principle and used it to develop a synergy measure (Globerson et al. 2009). Also, Griffith and Koch recently introduced a new synergy measure (Griffith and Koch 2012). Unlike the measures discussed so far, these additional methods are iterative and do not, to our knowledge, produce closed form solutions. Given their iterative nature, these measures may prove more computationally costly in comparison to the previously discussed information measures, especially when examining the interactions between many variables. For the sake of simplicity, we will not apply these methods to the following example systems.

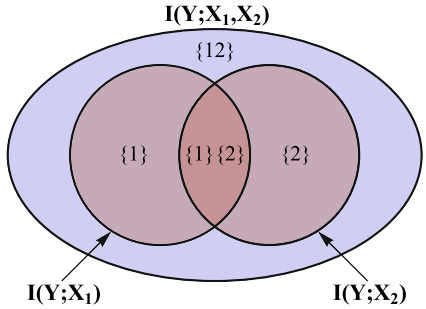

Fig. 1 Two variable partial information diagram. Figure taken from work by Williams and Beer with the authors' permission (Williams and Beer 2010)

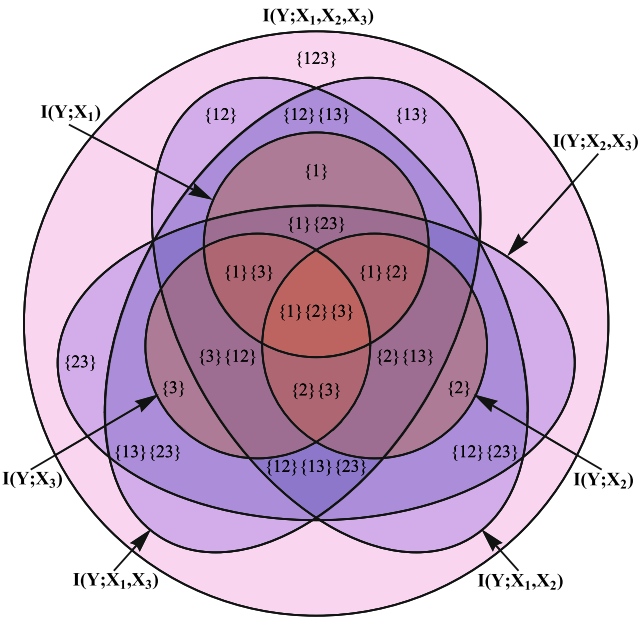

Fig. 2 Three variable partial information diagram. Figure taken from work by Williams and Beer with the authors' permission (Williams and Beer 2010)

Besides these methods, information geometry (Amari 2001) has also been used to examine neural networks (Ohiorhenuan and Victor 2011), as have various maximum entropy methods (Schneidman et al. 2006; Shlens et al. 2006; Tang et al. 2008). Finally, L. Martignon et al. discussed several methods, including the interaction information and methods similar to the maximum entropy methods discussed above to examine higher-order correlations in neural systems (Martignon et al. 2000). Though these methods also employ information theory, they are substantially different from the information theoretic analyses that can be performed with the measures discussed so far. Briefly, maximum entropy approaches can be used to indicate when a model based on pairwise correlations is insufficient to fit the probability distribution of network-wide correlated states (Yu et al. 2011; Fairhall et al. 2012; Shimazaki et al. 2012). Although information theory is often used to quantify the quality of this fit, the interactions involving three or more neurons are not in general quantified in terms of synergy, redundancy, etc. In comparison, the information measures discussed so far seek to directly measure interactions of varying orders on a variable-group by variable-group basis.

So, we will not discuss these alternative methods, but the reader is encouraged to explore them to see if they would better address his or her experimental question.

In addition to the information measures and information analysis techniques discussed so far, Schreiber introduced the transfer entropy to measure the information transfer between quantities that vary through time (Schreiber 2000). Because transfer entropy is focused on time series analysis and relationships between two variables, we have chosen not to apply it to the following example systems. It should be noted that the transfer entropy can be expressed in terms of the partial information decomposition (Williams and Beer 2011), transfer entropy has been applied to many systems (Marschinski and Kantz 2002; Lungarella and Sporn 2006; Honey et al. 2007; Ito et al. 2011; Lizier et al. 2011; Vicente et al. 2011), and that transfer entropy better captures information transfer in time series compared to time lagged mutual information and other methods (Schreiber 2000; Garofalo et al. 2009).

5 Example systems

We will now apply the multivariate information measures discussed above to several simple systems in an attempt to understand their similarities, differences, and uses. These systems have been chosen to maximize the contrast between the information measures, but many other systems exist for which the information measures produce identical results.

5.1 Examples 1-3: two-input Boolean logic gates

The first set of examples we will consider are simple Boolean logic gates. These logic gates are well known across many disciplines and offer a great deal of simplicity. The results presented in Table 1 highlight some of the commonalities and disparities between the various information measures. It should be noted that, due to the simple structure of the Boolean logic gates, the total correlation is equal to the mutual information. Also, due to the fact that only two-input Boolean logic gates are being considered, the redundancy-synergy index and Varadan's synergy are equal to the interaction information. Additional examples will highlight differences between these information measures.

All of the information measures provide a similar result for Example 1 (XOR-gate) (with the exception of the dual total correlation, see below). The interaction information, ΔI, the redundancy-synergy index, Varadan's synergy, and the partial information decomposition all indicate that the entire bit of information between Y and $\{X_1, X_2\}$ is accounted for by synergy. We might expect this result because, to know the state of Y for an XOR-gate, the state of both X_1 and X_2 must be known.

Table 1 Examples 1 to 3: two-input Boolean logic gates

			Ex. 1 XOR	Ex. 2 X_1	Ex. 3 AND
$p(x_1, x_2, y)$	x_1	x_2	y	y	y
1/4	0	0	0	0	0
1/4	1	0	1	1	0
1/4	0	1	1	0	0
1/4	1	1	0	1	1
$I(X_1; Y)$			0	1	0.311
$I(X_2; Y)$			0	0	0.311
$I(X_1, X_2; Y)$			1	1	0.811
$II(X_1; X_2; Y)$			1	0	0.189
$TC(X_1; X_2; Y)$			1	1	0.811
$DTC(X_1; X_2; Y)$			2	1	1
$\Delta I(X_1, X_2; Y)$			1	0	0.104
$RSI(X_1, X_2; Y)$			1	0	0.189
$VS(X_1, X_2; Y)$			1	0	0.189
$\Pi_R(Y; \{1\}\{2\})$			0	0	0.311
$\Pi_R(Y; \{1\})$			0	1	0
$\Pi_R(Y; \{2\})$			0	0	0
$\Pi_R(Y; \{12\})$			1	0	0.5

XOR-gate: all information measures produce results that indicate the presence of synergy. X_1-gate: the partial information decomposition succinctly identifies a relationship between X_1 and Y. AND-gate: the partial information decomposition identifies both synergistic and redundant interactions. The interaction information finds only a synergistic interaction. ΔI identifies the importance of correlations between X_1 and X_2

The results for Example 2 (X_1-gate) demonstrate the potential utility of the partial information decomposition. The unique information term from X_1 is equal to one bit, thus indicating that the X_1 variable entirely and solely determines the state of the output variable. This result is confirmed by the truth-table. This result can also be seen by considering the values of the other measures together (for instance, the three mutual information measures), but the partial information decomposition provides these results more succinctly.

More significant differences among the information measures appear when considering Example 3 (AND-gate). The partial information decomposition produces the result that 0.311 bits of information are provided redundantly and 0.5 bits are provided synergistically. Since each X variable provides the same amount of information about each state of Y (see Eq. (30)), the partial information decomposition finds that all of the mutual information between each X variable individually and the Y variable is redundant. As a result of this, no information is provided uniquely, and subsequently, the entirety of the remaining 0.5 bits of information between

Y and $\{X_1, X_2\}$ must be synergistic. From this, we can see in action the fact that the partial information decomposition emphasizes the *amount* of information that each X variable provides about each state of Y *considered individually*.

The interaction information, and by extension the redundancy-synergy index and Varadan's synergy, are limited to returning only a synergy value of 0.189 bits for the AND-gate. This value is produced because the mutual information between Y and $\{X_1, X_2\}$ contains an excess of 0.189 bits beyond the sum of the mutual informations between each X variable individually and the Y variable. So, here we can see in action the interpretation of the interaction information as the *amount* of information provided by the X variables taken together about Y, beyond what they provide individually. Also, the AND-gate allows us to see the relationship between the interaction information and the partial information decomposition as expressed by Eq. (35).

The value of ΔI for the AND-gate can be elucidated by examining the values of the conditional probability distributions that are relevant to the calculation of ΔI (Table 2). From these results, it is clear that if we use the independent model, and we are presented with the state $x_1 = 1$ and $x_2 = 1$, we would conclude that there is a one-quarter chance that $y = 0$ and a three-quarters chance that $y = 1$. If we use the actual data, then we know that, for that specific state, y must equal 1. This example points to a subtle, but critical difference between ΔI and the other multivariate information measures. Namely, the other information measures are concerned with discerning the interactions among the variables in the situation where you know the values of all the variables simultaneously. On the other hand, ΔI is concerned with *comparing* that situation to one where only the probability distributions of the variable pairs Y and X_i are known and the X_i variables are assumed to be independent (i.e. the independent model described by Eq. (19) (see Section 5.2 for further discussion of this topic).

The values of the dual total correlation for the XOR-gate example in Table 1 demonstrate a crucial difference between the dual total correlation and the other multivariate information measures. Namely, the dual total correlation does not differentiate between the X and Y variables. So, dependencies between all variables are treated equally. In the case of the XOR-gate, the entropy of any variable conditioned on the other two is zero. However, the joint entropy between all variables is 2 bits, so the dual total correlation is equal to 2 bits. Clearly, this result is greater than $I(X_1, X_2; Y)$ for this example. So, if we assume the synergy and redundancy are some portion of $I(X_1, X_2; Y)$, the dual total correlation cannot be the synergy or the redundancy. However, this result is not surprising given the fact that, if we assume the synergy and redundancy are some portion of $I(X_1, X_2; Y)$, the synergy and redundancy require some differentiation between the X variables and Y variables. Since the dual total correlation does not incorporate this distinction, we should expect that it measures a fundamentally different quantity (see Section 5.4 for further discussion of this topic).

Table 2 Values of conditional probabilities used to calculate ΔI for the AND-gate

y	x_1	x_2	$p_{ind}(y \mid x_1, x_2)$	$p(y \mid x_1, x_2)$
0	0	0	1	1
0	1	0	1	1
0	0	1	1	1
0	1	1	0.25	0
1	0	0	0	0
1	1	0	0	0
1	0	1	0	0
1	1	1	0.75	1

5.2 Example 4: ΔI is not bound by I

Another relevant example for ΔI is shown in Table 3. The crucial point to draw from this example is that ΔI can be greater than $I(X_1, X_2; Y)$. This appears to be in conflict with the intuitive notion of synergy as some part of the information the X variables provide about the Y variable. Why, in this case, ΔI is greater than $I(X_1, X_2; Y)$ is not

Table 3 Example 4. For this system, $\Delta I(X_1, X_2; Y)$ is greater than $I(X_1, X_2; Y)$. Schneidman et. al. also present an example that demonstrates that $\Delta I(X_1, X_2; Y)$ is not bound by $I(X_1, X_2; Y)$ (Schneidman et al. 2003a)

			Ex. 4
$p(x_1, x_2, y)$	x_1	x_2	y
1/10	0	0	0
1/10	1	1	0
2/10	0	0	1
6/10	1	1	1
$I(X_1; Y)$			0.0323
$I(X_2; Y)$			0.0323
$I(X_1, X_2; Y)$			0.0323
$II(X_1; X_2; Y)$			−0.0323
$TC(X_1; X_2; Y)$			0.9136
$DTC(X_1; X_2; Y)$			0.8813
$\Delta I(X_1, X_2; Y)$			0.0337
$RSI(X_1, X_2; Y)$			−0.0323
$VS(X_1, X_2; Y)$			−0.0323
$\Pi_R(Y; \{1\}\{2\})$			0.0323
$\Pi_R(Y; \{1\})$			0
$\Pi_R(Y; \{2\})$			0
$\Pi_R(Y; \{12\})$			0

immediately clear. To better understand this result, we can examine the difference between ΔI and $I(X_1, X_2; Y)$. Using Eqs. (5), (20), and (22), this difference can be expressed as:

$$I(S;Y) - \Delta I(S;Y) = \sum_{\vec{x}\in S, y\in Y} p(y,\vec{x}) \log\left(\frac{p_{ind}(\vec{x}|y)}{p_{ind}(\vec{x})}\right) \quad (37)$$

The quantity expressed on the RHS of Eq. (37), though similar in form, is not a mutual information. Based on the example in Table 3 and the examples in Table 1, this quantity can be positive or negative. Schneidman et. al. further explore this and other noteworthy features of ΔI (Schneidman et al. 2003a). Fundamentally, ΔI is a comparison between the complete data and an independent model (as expressed in Eq. (19)). As Schneidman et. al. note, alternative models could be chosen for the purpose of measuring the importance of correlations between the X variables in the data. Perhaps the best way to conceptualize the result from Table 3 is to note that in this case the information cost of assuming the X variables act independently of each other on Y (ΔI) is greater than the information cost of assuming Y is independent of the X variables (the mutual information). We wish to emphasize that ΔI can provide useful information about a system, but that it measures a fundamentally different quantity in comparison to the other multivariate information measures.

5.3 Example 5: amount vs. content

The example shown in Table 4 highlights some interesting differences between the information measures, especially regarding the partial information decomposition. Results from the partial information decomposition indicate that 1 bit of information about Y is provided redundantly by X_1 and X_2, while 1 bit is provided synergistically. This situation is similar to the AND-gate above. Each X variable provides 1 bit of information about Y, but both X variables provide the same amount of information about each state of Y. So, the partial information decomposition concludes that all of the information is redundant. It should be noted that this is the case despite the fact that X_1 *and* X_2 *provide information about different states of* Y. X_1 can differentiate between $y = 0$ and $y = 2$ on the one hand and $y = 1$ and $y = 3$ on the other, while X_2 can differentiate between $y = 0$ and $y = 1$ on the one hand and $y = 2$ and $y = 3$ on the other. Even though the X variables provide information about different states of Y, the partial information decomposition is blind to this distinction and concludes, since the X variables provide the same *amount* of information about each state of Y, that their contributions are redundant. Because all of the mutual information between each X variable considered individually is taken up by redundant information, the partial information decomposition concludes there is no unique information and, thus, the remaining 1 bit of information must be synergistic.

Table 4 Example 5: Y obtains a different state for each unique combination of X_1 and X_2

			Ex. 5
$p(x_1, x_2, y)$	x_1	x_2	y
1/4	0	0	0
1/4	1	0	1
1/4	0	1	2
1/4	1	1	3
$I(X_1; Y)$			1
$I(X_2; Y)$			1
$I(X_1, X_2; Y)$			2
$II(X_1; X_2; Y)$			0
$TC(X_1; X_2; Y)$			2
$DTC(X_1; X_2; Y)$			2
$\Delta I(X_1, X_2; Y)$			0
$RSI(X_1, X_2; Y)$			0
$VS(X_1, X_2; Y)$			0
$\Pi_R(Y; \{1\}\{2\})$			1
$\Pi_R(Y; \{1\})$			0
$\Pi_R(Y; \{2\})$			0
$\Pi_R(Y; \{12\})$			1

The partial information decomposition indicates the presence of redundancy because the X variables provide the same amount of information about each state of Y, despite the fact that the X variables provide information about different states of Y. The interaction information and ΔI provide null results. Griffith and Koch also discuss this example in relation to multivariate information measures (Griffith and Koch 2012). In this example, the Y variable can also be thought of as a joint variable $\{X_1, X_2\}$

Example 5 demonstrates the conditions for null results from the interaction information and ΔI. When considering the relationship between one of the X variables and Y, we see that knowing the state of the X variable reduces the uncertainty about Y by 1 bit in all cases. However, knowing both X variables only provides 2 bits of information about Y. So, the interaction information must be zero because no additional information about Y is gained or lost by knowing both X variables together compared to knowing them each individually. Similarly, ΔI must be zero because the knowledge of the state of X_1 and X_2 simultaneously does not provide any additional knowledge about Y compared to the independent models for the relationships between each X variable and Y.

5.4 Example 6: zero target entropy

The example shown in Table 5 demonstrates a significant feature of the total correlation. Even when no information is

Table 5 Example 6. All information measures, with the exceptions of the total correlation and the dual total correlation, are zero

			Ex. 6
$p(x_1, x_2, y)$	x_1	x_2	y
1/2	0	0	0
1/2	1	1	0
$I(X_1; Y)$			0
$I(X_2; Y)$			0
$I(X_1, X_2; Y)$			0
$II(X_1; X_2; Y)$			0
$TC(X_1; X_2; Y)$			1
$DTC(X_1; X_2; Y)$			1
$\Delta I(X_1, X_2; Y)$			0
$RSI(X_1, X_2; Y)$			0
$VS(X_1, X_2; Y)$			0
$\Pi_R(Y; \{1\}\{2\})$			0
$\Pi_R(Y; \{1\})$			0
$\Pi_R(Y; \{2\})$			0
$\Pi_R(Y; \{12\})$			0

The total correlation and the dual total correlation produce non-zero results because they detect interactions between the X variables, while the other measures that differentiate between the X variables and the Y variable are zero

passing to one of the variables considered, the total correlation and the dual total correlation can still produce non-zero results if interactions are present between other variables in the system. This result can be clearly understood using the expression for the total correlation in Eq. (15) and the expression for the dual total correlation in Eq. (18). The total correlation sums the information passing between variables from the smallest scale (two variables) to the largest scale (n variables). It will detect relationships at all levels and it is unable to differentiate between those levels. The dual total correlation compares the total correlation to the amount of information passing between each individual variable and all other variables considered together as a single vector valued variable. As with the dual total correlation, the total correlation does not differentiate between the X and Y variables, unlike several of the other information measures.

In this case, Y has no entropy, so all information measures that describe interactions between the X variables and the target variable Y (i.e., all of the other information measures considered here) are zero. This is expected since all of the other information measures are either explicitly focused on the relationship between the X variables and the Y variable or only focus on interactions that involve all variables.

5.5 Examples 7 and 8: three-input Boolean logic gates

The three-input Boolean logic gate examples shown in Table 6 allow for a comparison between the interaction information, the redundancy-synergy index, Varadan's synergy, and the partial information decomposition. Example 7 (three-way XOR-gate) produces similar results to the XOR-gate shown in Table 1. All of the information measures indicate the presence of a synergistic interaction or an interaction where the correlations between the X variables provide additional information. The partial information decomposition is able to localize the synergy to an interaction between all three X variables.

Significant differences appear between the information measures when an extraneous X_3 variable is added to a basic XOR-gate between X_1 and X_2 (Example 8, X_1X_2XOR-gate). In this case, the interaction information is zero because there is no synergy present between all three X

Table 6 Examples 7 and 8: three-input Boolean logic gates

				Ex. 7 3XOR	Ex. 8 X_1X_2XOR
$p(x_1, x_2, y)$	x_1	x_2	x_3	y	y
1/8	0	0	0	0	0
1/8	1	0	0	1	1
1/8	0	1	0	1	1
1/8	1	1	0	0	0
1/8	0	0	1	1	0
1/8	1	0	1	0	1
1/8	0	1	1	0	1
1/8	1	1	1	1	0
$I(X_1; Y)$				0	0
$I(X_2; Y)$				0	0
$I(X_3; Y)$				0	0
$I(X_1, X_2, X_3; Y)$				1	1
$II(X_1; X_2; X_3; Y)$				1	0
$TC(X_1; X_2; X_3; Y)$				1	1
$DTC(X_1; X_2; X_3; Y)$				3	2
$\Delta I(X_1, X_2, X_3; Y)$				1	1
$RSI(X_1, X_2, X_3; Y)$				1	1
$VS(X_1, X_2, X_3; Y)$				1	0
$\Pi_R(Y; \{12\})$				0	1
$\Pi_R(Y; \{123\})$				1	0

All partial information decomposition terms not shown in the table are zero. 3XOR: Three-way XOR-gate. All information measures produce consistent results. X_1X_2XOR: XOR-gate involving only X_1 and X_2. The redundancy-synergy index identifies a synergistic interaction and ΔI identifies the importance of correlations between the X variables. The partial information decomposition also identifies the variables involved in the synergistic interaction. The interaction information and Varadan's synergy do not identify a synergistic interaction

variables and the Y variable. This is despite the fact that the interaction information indicated synergy was present for the two X variable XOR-gate. Thus, we can see that the interaction information focuses only on interactions between *all* of the X variables and the Y variable. A similar result is observed with Varadan's synergy. Despite the fact that it indicated the presence of synergy in the two X variable XOR-gate, Varadan's synergy does not indicate synergy is present in this logic gate because it also focuses only on interactions between all of the X variables and the Y variable. Both the redundancy-synergy index and the partial information decomposition return results that indicate the presence of synergy between the X variables and the Y variable, but only the partial information decomposition is able to localize the synergy to the X_1 and X_2 variables.

5.6 Examples 9 to 13: simple model networks

In an effort to discuss results more directly applicable to several research topics, we will now apply the multivariate information measures to several variations of a simple model network. The general structure of the network is shown in Fig. 3. The network contains three nodes, each of which can be in one of two states (0 or 1) at any given point in time. The default state of each node is 0. At each time step, there is a certain probability, call it p_r, that a given node will be in state 1. The probability that a given node is in state 1 can also be increased if it receives a connection from another node. This driving effect is noted by p_{1y} for the connection from X_1 to Y, p_{12} for the connection from X_1 to X_2, and p_{2y} for the connection from X_2 to Y. All states of the network are determined simultaneously and are independent of the previous states of the network. (See Appendix C for further details regarding this model.)

For this simple system, we will discuss five simple combinations of p_r, p_{1y}, p_{12}, and p_{2y}. We chose to present these combinations of probabilities because they span interesting and often times difficult to analyze topologies. Also, we chose to present combinations with small probability values because many experiments in neuroscience are focused on events that happen relatively infrequently. However, the system can be easily modified to include any possible combinations of probabilities. The information theoretic results for these examples are presented in Table 7.

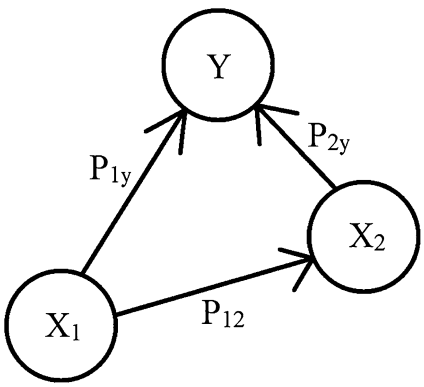

Fig. 3 Structure of model network used for Examples 9 to 13

Example 9 represents a system where the X nodes independently drive the Y node. Similarly to Example 5, the partial information decomposition indicates that the information from X_1 and X_2 is entirely redundant and synergistic. This result is somewhat counter intuitive because the X nodes act independently. Again, this is due to the structure of the minimum information in Eq. (30). Each X variable provides the same amount of information about each state of Y, so the partial information decomposition returns the result that all of the information provided by each X variable about Y is redundant. The interaction information returns a result that indicates the presence of synergy, though the magnitude of this interaction is less than the magnitudes of the synergy and redundancy results from the partial information decomposition. Note that this is the only network for which the interaction information indicates the presence of synergy.

Example 10 is similar to Example 9 with the exception that X_1 now also drives X_2. Several interesting results are produced for this example. For instance, the total correlation and dual total correlation are significantly elevated in comparison to the other examples. In this example, there is

Table 7 Examples 9 to 13: simple model network. All information values are in millibits

Diagram	Ex. 9	Ex. 10	Ex. 11	Ex. 12	Ex. 13
p_r	0.02	0.02	0.02	0.02	0.02
p_{12}	0	0.1	0.1	0.1	0.1
p_{1y}	0.1	0.1	0	0.1	0
p_{2y}	0.1	0.1	0.1	0	0
$I(X_1; Y)$	3.061	3.498	0.053	3.225	0
$I(X_2; Y)$	3.061	3.801	3.527	0.050	0
$I(X_1, X_2; Y)$	6.239	6.750	3.527	3.225	0
$II(X_1; X_2; Y)$	0.117	−0.548	−0.053	−0.050	0
$TC(X_1; X_2; Y)$	6.239	9.975	6.752	6.450	3.225
$DTC(X_1; X_2; Y)$	6.356	9.427	6.698	6.400	3.225
$\Delta I(X_1, X_2; Y)$	0.080	0.499	0.064	0.059	0
$RSI(X_1, X_2; Y)$	0.117	−0.548	−0.053	−0.050	0
$VS(X_1, X_2; Y)$	0.117	−0.548	−0.053	−0.050	0
$\Pi_R(Y; \{1\}\{2\})$	3.061	3.498	0.053	0.050	0
$\Pi_R(Y; \{1\})$	0	0	0	3.175	0
$\Pi_R(Y; \{2\})$	0	0.303	3.473	0	0
$\Pi_R(Y; \{12\})$	3.178	2.950	0	0	0

the maximum amount of interactions between all nodes. So, this result agrees with expectations because the total correlation and dual total correlation reflect the total amount of interactions at all scales between all variables. Also, ΔI obtains its highest value for this example because the actual data and the independent model from Eq. (19) that is used in the calculation of ΔI are more dissimilar due to the interactions between X_1 and X_2. For Example 9, ΔI has a lower value because there are no interactions between X_1 and X_2, so the independent model from Eq. (19) that is used in the calculation of ΔI is more similar to the actual data. Interestingly, the partial information decomposition does not indicate the presence of unique information from X_1, despite the fact that X_1 is directly influencing Y. As with example 9 above, the partial information decomposition returns the result that all of the information X_1 provides about Y is redundant. Thus, based on the structure of the minimum information function, there is no state of Y for which X_1 provides more information than X_2. In this case, this result is more intuitive because X_1 also drives X_2. The interaction information returns a significantly larger magnitude result for this example. This is intuitive given the fact that X_1 is driving X_2 and that both X variables are driving the Y variable. However, it should be noted that the magnitude of the interaction information is significantly less than the magnitude of the synergy and redundancy from the partial information decomposition. Also, the interaction information result implies the presence of redundancy, unlike Example 9.

Example 11 represents a common problem case when attempting to infer connectivity based solely on node activity. Node X_1 drives X_2, which in turn drives Y. If the activity of X_2 is not known, it would appear that X_1 is driving Y directly. The partial information decomposition returns the result that any information provided by X_1 about Y is redundant and that the vast majority of the information provided by X_1 and X_2 about Y is unique information from X_2. Both of these results appear to accurately reflect the structure of the network.

Example 12 also represents a common problem case when determining connectivity. Node X_1 drives Y and X_2. If the activity of X_1 is not known, it would appear that X_2 is driving Y, when, in fact, no connection exists from X_2 to Y. Similarly to Example 11, the partial information decomposition identifies the majority of the information from X_1 and X_2 about Y as unique information from X_1 and the remaining information as redundant. Again, these results appear to accurately reflect the structure of the network. However, it should be noted that these results are numerically similar to the results from Example 11 with X_1 and X_2 interchanged. Thus, it may be the case that the partial information decomposition may not be able to differentiate between situations like Examples 11 and 12 in some scenarios.

The final example (13) is similar to Example 6 above. In this case, no connections exist from X_1 or X_2 to Y, but X_1 drives X_2. Almost all of information measures indicate a lack of information transmission. However, the total correlation and the dual total correlation pick up the interaction between X_1 and X_2. The values of the total correlation and the dual total correlation increase as the number of connections in the model increase. This, again, demonstrates the fact that the total correlation and the dual total correlation measure interactions between all variables at all scales.

5.7 Analysis of dissociated neural culture

We will now present the results of applying the information measures discussed above to spiking data from a dissociated neural culture. We wish to emphasize that we were not attempting to answer any specific experimental question. Rather, we simply hoped to illustrate one type of analysis that is possible using these information measures.

The data we chose to analyze are described in Wagenaar et al. and are freely available online (Wagenaar et al. 2006a). The data contain multiunit spiking activity for each of 60 electrodes in the multielectrode array on which the dissociated neural culture was grown. Specifically, we used data from neural culture 2-2. All details regarding the production and maintenance of the culture can be found in (Wagenaar et al. 2006a). We analyzed recordings from eight points in the development in the culture: days *in vitro* (DIV) 4, 7, 12, 16, 20, 25, 31, and 33. The DIV 16 recording was 60 minutes long, while all others were 45 minutes long.

For this analysis, the data were binned at 16 ms. The probability distributions necessary for the computation of the information measures were created by examining the spike trains for groups of three non-identical electrodes. For a given group of electrodes, one electrode was labeled the Y electrode, while the other two were labeled the X_1 and X_2 electrodes. Then, for all time steps in the spike trains, the states of the electrodes (spiking or not spiking) were recorded at time t for the X_1 and X_2 electrodes and at time $t+1$ for the Y electrode. Next, by counting how many times each state appeared throughout the spike train, the joint probabilities $p(y_{t+1}, x_{1,t}, x_{2,t})$ were calculated, which were then used to calculate the information measures discussed above. This process was repeated for each group of non-identical electrodes. However, to avoid double counting, groups with swapped X variable assignments were only analyzed once. For instance, the group X_1 = electrode 3, X_2 = electrode 4, and Y = electrode 5 was analyzed, but the group X_1 = electrode 4, X_2 = electrode 3, and Y = electrode 5 was not analyzed. In order to compensate for the changing firing rate through development of the cultures, all information values for a given group were normalized by the entropy of the Y electrode.

To illustrate the statistical significance of the information measure values, we also created and analyzed a randomized data set from the original neural culture data for each DIV. The randomization was accomplished by splitting each electrode spike train at a randomly chosen point and swapping the two remaining pieces. By doing this, the structure of the electrode spike train is almost entirely preserved, but the temporal relationship between the electrode spike trains is significantly disrupted. We wish to emphasize that several alternative randomization schemes exist (spike swapping, spike jittering, inter-spike interval shuffling, etc.) and have been widely used in neuroscience, including applications outside information theory (Louie and Wilson 2001; Hatsopoulos et al. 2003; Beggs and Plenz 2004; Ikegaya et al. 2004; Rivlin-Etzion et al. 2006; Butts et al. 2007; Madhavan et al. 2007; Rolston et al. 2007; Wang et al. 2007; Fujisawa et al. 2008; Pazienti et al. 2008). Furthermore, it is probably the case that each information measure will behave differently under different randomization schemes. Also, the effectiveness of a randomization method is probably dependent on whether the information measure treats all variables equally or if it measures the interactions between a group of variables and another target variable. For instance, for $I(X_1, X_2; Y)$ it would be better to shift the X variables together to preserve their joint probability distribution and only disrupt the temporal

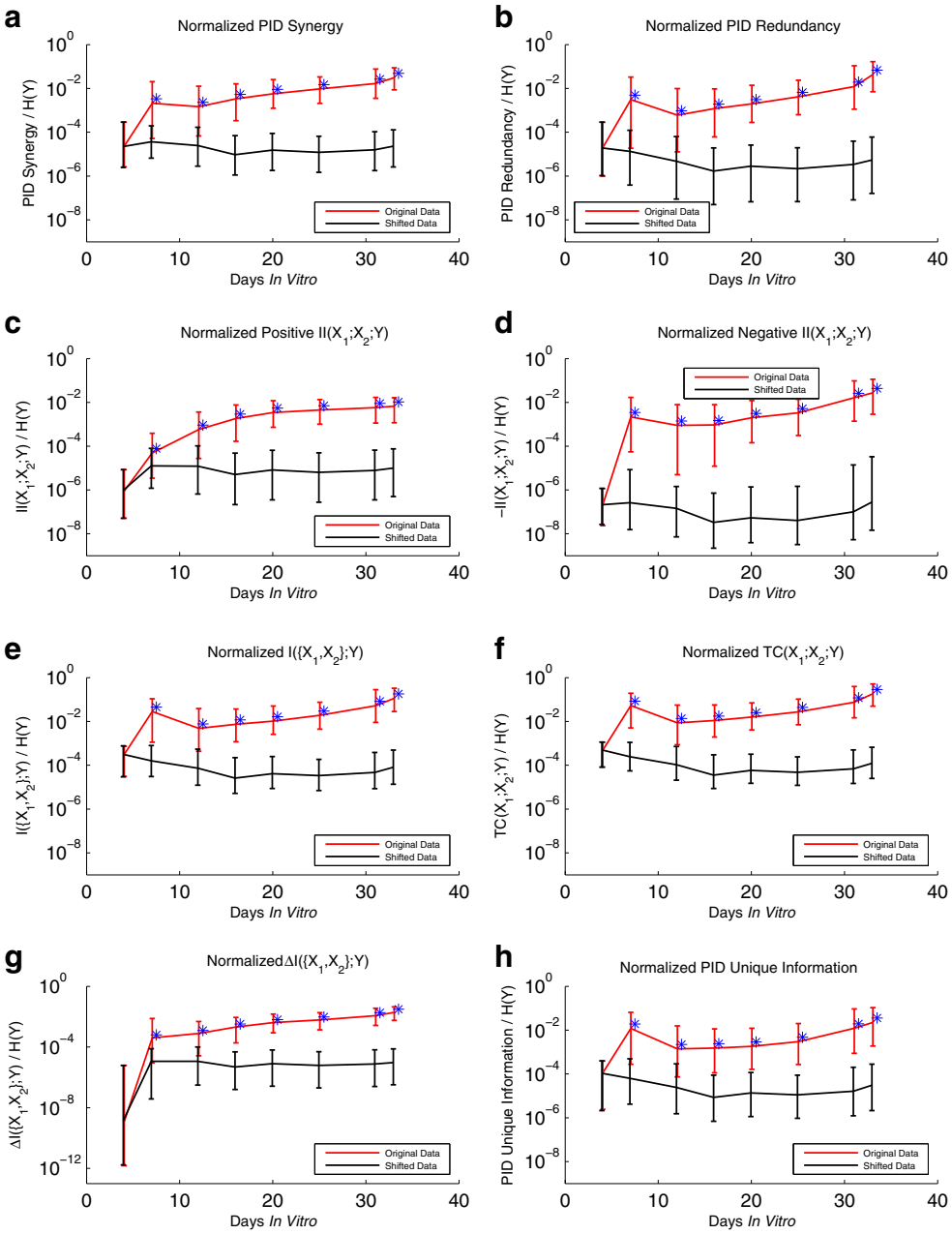

Fig. 4 Many information measures show changes over neural development. **a** PID Synergy. **b** PID Redundancy. **c** Positive Interaction Information values. **d** Negative Interaction Information values. Note: a group of electrodes can only appear in C or D, but not both. **e** Mutual Information. **f** Total Correlation. **g** ΔI. **h** PID Unique Information. All information values are normalized by the entropy of the Y electrode. The line plots show the 90th percentile, median, and 10th percentile of all the data for a given DIV. KS test results with $p < 0.01$ are marked with a blue asterisk. Note that as the culture matured, the total amount of information transmitted increased and the types of interactions present in the network changed

relationship between the X variables and the Y variable. Similar allowances should be made for the other measures, though this process may become significantly more complicated for other measures. To our knowledge, no one has thoroughly researched the effectiveness of these different randomization methods for the information measures discussed herein. It is beyond the scope of this review to do so, thus, for this illustration, we have chosen to use one randomization method for all information measures for the sake of simplicity. However, researchers are cautioned to carefully consider the effect different randomization methods may have on different information measures when performing a more rigorous comparison between information measures applied to experimental data.

Following the randomization of the data for each DIV, for each information measure, we logarithmically binned 5000 randomly chosen non-zero information values from all possible triplets of electrodes to obtain information value distributions for the original data and for the shuffled data. Forty bins were used to span the non-zero data, with the same bins being used for both data sets. A Kolmogorov-Smirnov (KS) test was then performed on the distributions to assess the statistical significance of the information results.

The results of these analyses are presented in Figs. 4 and 5. The results shown in Fig. 4 indicate that at day 4 essentially no information was being transmitted in the network because all of the information values were not significantly different from the randomized data. However, by day 7, a great deal of information was being transmitted, as can be seen by the peaks in mutual information (Fig. 4e) and the total correlation (Fig. 4f). As the culture continued to develop after day 7, most information measures decreased and then slowly increased to maxima on the last day, DIV 33. Interestingly, ΔI (Fig. 4f) showed an increase at day 7, but then a steady increase afterwards. The total correlation (Fig. 4f) mimics the changes in the mutual information (Fig. 4e), but because the total correlation measures the total amount of information being transmitted among the X and Y variables, it possessed higher values than the mutual information.

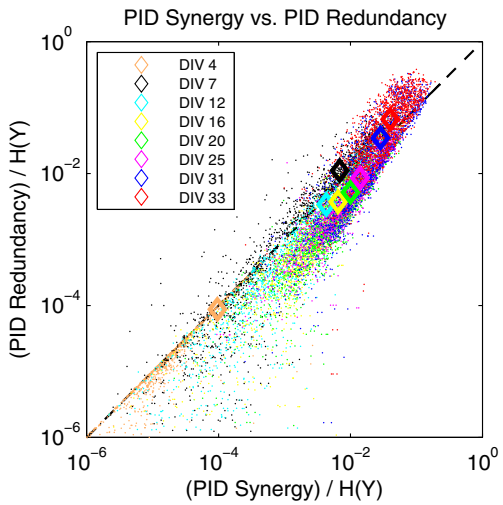

Fig. 5 The balance of PID synergy and PID redundancy changed during development. Each data point represents the information values for one group of electrodes (only 2 % of the data are shown to improve clarity). Diamonds represent mean values for a given DIV for all groups of electrodes

The relationship between the interaction information and the partial information decomposition was also illustrated through development. As the culture developed, the PID synergy was larger than the PID redundancy. Then, between days 31 and 33, the PID synergy became significantly smaller than the PID redundancy. In the interaction information, this relationship was expressed by positive values through most of the culture's development, with the exception of large negative values at days 7 and 33. However, notice that groups of electrodes with positive and negative interaction information values were found in each recording. To further investigate this relationship, we plotted the distribution of PID synergy and PID redundancy for groups of electrodes (Fig. 5). This plot shows that the network contained groups of electrodes with slightly more PID redundancy than PID synergy at day 7, but that, after that point, the total amount of information decreased and became more biased towards PID synergy at day 12. From that point, the total amount of information increased up to the last recording where the network was once again biased towards PID redundancy. So, we can relate the results from the partial information decomposition and the interaction information using Fig. 5 by noting that, while the PID synergy and PID redundancy for a given group of electrodes determines a points position in Fig. 5, the interaction information describes how far that point is from the equilibrium line. Given the fact that many points in Fig. 5 are near the equilibrium line, the partial information decomposition finds that many groups of electrodes contain synergistic and redundant interactions simultaneously. This feature would be lost by only examining the interaction information.

Obviously, this analysis could be made significantly more complex and interesting. For instance, the analysis could be improved by including more data sets, varying the variable assignments, using different bin sizes, using more robust methods to test statistical significance, and so forth. However, based on this simple illustration, we believe that it is clear that the information analysis methods discussed herein could be used to address interesting questions in this or other systems. For instance, it may be possible to relate these changes through development to previous work on changes in dissociated cultures through development (Kamioka et al. 1996; Wagenaar et al. 2006a, 2006b; Pasquale et al.2008; Tetzlaff et al.2010). Also, these information measures could be used

to study the changes in synergy and redundancy through development, as well as changes through development in the amount of information being transferred among different numbers of neurons.

6 Discussion

Based on the results from several simple systems, we were able to explore the properties of the multivariate information measures discussed in this paper. We will now discuss each measure in turn.

The oldest multivariate information measure - the interaction information - was shown to focus on interactions between all X variables and the Y variable using the three-input Boolean logic gate examples. Furthermore, the two-input AND-gate demonstrated how the interaction information is related to the excess information provided by both X variables about the Y variable beyond the total amount of information those X variables provide about Y when considered individually. Also, that example demonstrated the relationship between the interaction information and the partial information decomposition as shown in Eq. (35). For the model network examples, the interaction information had its largest magnitude when the interactions were present between all three nodes. Also, for these examples, the interaction information indicated the presence of synergy when both nodes X_1 and X_2 drove Y, but not each other (Example 9), while it indicated the presence of redundancy when either node X_1 or X_2 drove Y and X_1 drove Y (Examples 10 to 12). When the interaction information was applied to data from a developing neural culture, it showed changes in the type of interactions present in the network during development.

In contrast to the interaction information, the total correlation was shown to sum interactions among all variables at all scales using Example 6. In other words, the value of the total correlation for any system incorporates interactions between groups of variables at all scales. This feature was made apparent using the model network examples. There, the total correlation increased with the number of connections present in the network. Furthermore, the total correlation is symmetric with regard to all variables considered, whereas the other information measures focus on the relationship between the set of X variables and the Y variable. When applied to the data from the neural culture, the total correlation and the mutual information both showed increases in the total amount of information being transmitted in the network through development.

The dual total correlation was found to be similar to the total correlation in that both do not differentiate between the X and Y variables. Also, like the total correlation, the dual total correlation increased with the number of connections in the model network examples. The function of the dual total correlation was also highlighted with the XOR-gate example. There, we saw that the dual total correlation compares the uncertainty with regards to all variables to the total uncertainty that remains about each variable if all the other variables are known.

Using the AND-gate, ΔI was shown to measure a subtly different quantity compared to the other information measures. The other information measures seek to evaluate the interactions between the X variables and the Y variable given that one knows the values of all variables simultaneously (i.e. in the case that the total joint probability distribution is known). ΔI compares that situation to a model where it is assumed that the X variables act independently of one another in an effort to measure the importance of knowing the correlations between the X variables. Clearly, this goal is similar to the goals of the other information measures. However, given the fact that ΔI can be greater than $I(X_1, X_2; Y)$, as was shown in Example 4, and if we assume the synergy and redundancy are some portion of $I(X_1, X_2; Y)$, ΔI cannot be the synergy or the redundancy. ΔI can provide useful information about a system, but the distinction between the structure of ΔI and the other information measures, along with the fact that ΔI cannot be the synergy or redundancy as previously defined, should be considered when choosing the appropriate information measure with which to perform an analysis. Unlike several of the other information measures which showed changes in the types of interactions present in the developing neural culture, ΔI showed a general trend of increasing importance of correlations in the network throughout development.

The redundancy-synergy index and Varadan's synergy are identical to the interaction information when only two X variables are considered. However, when we examined three-input Boolean logic gates, we found that Varadan's synergy - like the interaction information - was unable to detect a synergistic interaction among a subset of the X variables and the Y variable. The redundancy-synergy index was able to detect this synergy, but it was unable to localize the subset of X variables involved in the interaction.

The partial information decomposition provided interesting and possibly useful results for several of the example systems. When applied to the Boolean logic gates, the partial information decomposition was able to identify the X variables involved in the interactions, unlike all other information measures. This is primarily due to the fact that the partial information decomposition produces several information terms to quantify different types of interactions simultaneously, unlike the other information measures that produce only one value. Using the AND-gate example, we saw that the partial information decomposition found that both synergy and redundancy were present in the system, unlike the interaction information, which indicated only

Table 8 Overview of the information measures

Name	Formula(s)	Variable Partition	Design Goal	Comments
Interaction Information (II)	Eq. (7)–(10)	All variables treated equally	Measure synergy and redundancy between S and Y	II only measures interactions between Y and all of S. Also, it does not allow for simultaneous redundancy and synergy.
Total Correlation (TC)	Eq. (13)–(15)	All variables treated equally	Measure the total information passing between all variables	
Dual Total Correlation (DTC)	Eq. (16)–(18)	All variables treaded equally	Measure the total information passing between each variable and all other variables beyond the TC	
ΔI	Eq. (22)	Two distinct groups of variables	Measure the importance of correlations between X variables in interactions with the Y variable	It is unclear how the results from ΔI relate to synergy (see Section 5.2)
Redundancy-Synergy Index (RSI)	Eq. (23)	Two distinct groups of variables	Measure the synergy and redundancy between S and Y	RSI does not allow for simultaneous redundancy and synergy.
Varadan's Synergy (VS)	Eq. (24)	Two distinct groups of variables	Measure the synergy and redundancy between S and Y	VS only measures interactions between Y and all of S. Also, VS does not allow for simultaneous redundancy and synergy.
Partial Information Decomposition (PID)	3 Variable Case: Eq. (31)–(34)	Two distinct groups of variables	Measure the synergy and redundancy between S and Y	PID allows for simultaneous synergy and redundancy, though its definition of synergy may be subtly different from the expected definition (see Section 5.3).

synergy was present. Perhaps the most illuminating example system for the partial information decomposition was Example 5. In that case, the partial information decomposition concluded that each X variable provided entirely redundant information because each X variable provided the same amount of information about each state of Y, even though each X variable provided information about different states of Y. This point highlights how the partial information decomposition defines redundancy via Eq. (30). It calculates the redundant contributions based only on the *amount* of information each X variable provides about each state of Y. In the developing neural culture, the partial information decomposition, similar to the interaction information, showed a changing balance between synergy and redundancy through development. However, unlike the interaction information, the partial information decomposition was able to separate simultaneous synergistic and redundant interactions.

Based on the results from the various example systems above and the theoretical structure of each measure, we have

Table 9 Overview of example application of the information measures in typical neuroscience experiments

	Example Experiment 1	Example Experiment 2	Example Experiment 3
General Topic	Encoding	Network Connectivity	Whole Network Behavior
Measurements	Record a visual stimulus and the neural activity of an animal	Record the spontaneous neural activity of an acute slice using a multi-electrode array	Record the spontaneous neural activity of an acute slice using a multi-electrode array
Y Variable	Stimulus variable	Activity of one neuron	None
X Variables	Activity of n neural sources	Activity of n neurons	Activity of n neurons
Experimental Question	How do the neural sources encode the stimulus?	How do the n neurons effectively determine activity of the single neuron?	How much information do the neurons share with each other?
Time Delay	X is time lagged from Y	Y is time lagged from X	None
Information Measures	II, ΔI, RSI, VS, or PID	II, ΔI, RSI, VS, or PID	TC or DTC

created a table that contains an overview of each measure (Table 8) and a table that contains information about when to use the various information measures in three example neuroscience experiments (Table 9).

7 Conclusion

We applied several multivariate information measures to simple example systems in an attempt to explore the properties of the information measures. We found that the information measures produce similar or identical results for some systems (e.g. XOR-gate), but that the measures produce different results for other systems. In examining these results, we found several subtle differences between the information measures that impacted the results. Then, we applied the information measures to spiking data from a neural culture through its development. Based on this illustrative analysis, we saw interesting changes in the amount of information being transmitted and the interactions present in the network.

We wish to emphasize that none of these information measures is the "right" measure. All of them produce results that can be used to learn something about the system being studied. We hope that this work will assist other researchers as they deliberate on the specific questions they wish to answer about a given system so that they may use the multivariate information measures that best suit their goals.

Conflict of Interest The authors declare that they have no conflict of interest.

Acknowledgments We would like to thank Paul Williams, Randy Beer, Alexander Murphy-Nakhnikian, Shinya Ito, Ben Nicholson, Emily Miller, Virgil Griffith, and Elizabeth Timme for providing useful comments. We would also like to thank the anonymous reviewers for their helpful comments on this paper. Their input during the revision process was invaluable.

Appendix A: Additional total correlation derivation

Equation (14) can be rewritten as Eq. (15) by adding and subtracting several joint entropy terms and then using Eq. (2). For instance, when $n = 3$, we have:

$$\begin{aligned} TC(S) &= \left(\sum_{X_i \in S} H(X_i)\right) - H(S) \\ &= H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) \\ &= H(X_1) + H(X_2) - H(X_1, X_2) + H(X_1, X_2) \\ &\quad + H(X_3) - H(X_1, X_2, X_3) \\ &= I(X_1; X_2) + I(X_1, X_2; X_3) \end{aligned} \tag{38}$$

A similar substitution can be peformed for $n > 3$.

Appendix B: Additional dual total correlation derivation

Equation (16) can be rewritten as Eq. (18) by substituting the expression for the total correlation in Eq. (14) and then applying Eq. (2).

$$\begin{aligned} DTC(S) &= \left(\sum_{X_i \in S} H(S/X_i)\right) - (n-1)H(S) \\ &= \left(\sum_{X_i \in S} H(S/X_i) + H(X_i)\right) - nH(S) - TC(S) \\ &= \left(\sum_{X_i \in S} I(S/X_i; X_i)\right) - TC(S) \end{aligned} \tag{39}$$

Appendix C: Model network

Given values for p_r, p_{1y}, p_{12}, and p_{2y}, the relevant conditional probabilities can be calculated in the following way:

$$p(x_1 = 1) = p_r \tag{40}$$

$$p(x_1 = 0) = 1 - p_r \tag{41}$$

$$p(x_2 = 1|x_1 = 1) = p_r + p_{12} - p_r p_{12} \tag{42}$$

$$p(x_2 = 0|x_1 = 1) = 1 - p(x_2 = 1|x_1 = 1) \tag{43}$$

$$p(x_2 = 1|x_1 = 0) = p_r \tag{44}$$

$$p(x_2 = 0|x_1 = 0) = 1 - p_r \tag{45}$$

$$p(y = 1|x_1 = 0, x_2 = 0) = p_r \tag{46}$$

$$p(y = 0|x_1 = 0, x_2 = 0) = 1 - p_r \tag{47}$$

$$p(y = 1|x_1 = 1, x_2 = 0) = p_r + p_{1y} - p_r p_{1y} \tag{48}$$

$$p(y=0|x_1=1, x_2=0)=1-p(y=1|x_1 = 1, x_2 = 0) \tag{49}$$

$$p(y = 1|x_1 = 0, x_2 = 1) = p_r + p_{2y} - p_r p_{2y} \tag{50}$$

$$p(y=0|x_1=0, x_2=1)=1-p(y=1|x_1=0, x_2=1) \tag{51}$$

$$\begin{aligned} p(y=1|x_1=1, x_2=1) &= p_r+p_{1y} + p_{2y} - p_r p_{1y} - p_r p_{2y} \\ &\quad -p_{1y} p_{2y}+p_r p_{1y} p_{2y} \end{aligned} \tag{52}$$

$$p(y=0|x_1=1, x_2=1)=1-p(y\ =\ 1|x_1=1, x_2=1) \tag{53}$$

Table 10 Joint probabilities for examples 9 to 13

Diagram			Ex. 9	Ex. 10	Ex. 11	Ex. 12	Ex. 13
							
p_r			0.02	0.02	0.02	0.02	0.02
p_{12}			0	0.1	0.1	0.1	0.1
p_{1y}			0.1	0.1	0	0.1	0
p_{2y}			0.1	0.1	0.1	0	0
x_1	x_2	y			$p(x_1, x_2, y)$		
0	0	0	0.9412	0.9412	0.9412	0.9412	0.9412
1	0	0	0.0173	0.0156	0.0173	0.0156	0.0173
0	1	0	0.0173	0.0173	0.0173	0.0192	0.0192
1	1	0	0.0003	0.0019	0.0021	0.0021	0.0023
0	0	1	0.0192	0.0192	0.0192	0.0192	0.0192
1	0	1	0.0023	0.0021	0.0004	0.0021	0.0004
0	1	1	0.0023	0.0023	0.0023	0.0004	0.0004
1	1	1	0.0001	0.0005	0.0003	0.0003	0.0000

Once these conditional probabilities are defined, the joint probabilities $p(y, x_1, x_2)$ can be calculated using the general relationship between joint and conditional probabilities:

$$p(A = a, B = b) = p(A = a|B = b)p(B = b) \tag{54}$$

The joint probabilities for the examples discussed in the main text of the article are shown in Table 10.

References

Abdallah, S.A., & Plumbley, M.D. (2010). A measure of statistical complexity based on predictive information. arXiv:1012.1890v1.

Amari, S.I. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, *8*(9), 1379.

Amari, S. (2001). *IEEE Transactions on Information Theory*, *47*, 1701.

Anastassiou, D. (2007). *Molecular Systems Biology*, *3*, 83.

Averbeck, B.B., Latham, P.E., Pouget, A. (2006). *Nature Reviews Neuroscience*, *7*, 358.

Beggs, J.M., & Plenz, D. (2004). Neuronal avalanches are diverse and precise activity patterns that are stable for many hours in cortical slice cultures. *Journal of Neuroscience*, *24*(22), 5216.

Bell, A.J. (2003). *International workshop on independent component analysis and blind signal separation,* (p. 921).

Berrou, C., Glavieux, A., Thitimajshima, P. (1993). In *Proceedings of IEEE International Conference on Communications* (Vol. 2, p. 1064).

Bettencourt, L.M.A., Stephens, G.J., Ham, M.I., Gross, G.W. (2007). *Physical Review E*, *75*, 021915.

Bettencourt, L.M.A., Gintautas, V., Ham, M.I. (2008). *Physical Review Letters*, *100*, 238701.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R.R., Warland, D. (1991). *Science*, *252*, 1854.

Borst, A., & Theunissen, F.E. (1999). *Nature Neuorscience*, *2*, 947.

Brenner, N., Strong, S.P., Koberle, R., Bialek, W., de Ruyter van Steveninck, R.R. (2000). *Neural Computation*, *12*, 1531.

Butte, A.J., & Kohane, I.S. (2000). In *Pacific Symposium on Biocomputing* (Vol. 5, p. 415).

Butts, D.A., & Rokhsar, D.S. (2001). *Journal of Neuroscience*, *21*, 961.

Butts, D.A., Weng, C., Jin, J., Yeh, C.I., Lesica, N.A., Alonso, J.M., Stanley, G.B. (2007). *Nature Letters*, *449*, 92.

Cerf, N.J., & Adami, C. (1997). *Physical Review A*, *55*, 3371.

Chanda, P., Zhang, A., Brazeau, D., Sucheston, L., Freudenheim, J.L., Ambrosone, C., Ramanathan, M. (2007). *American Journal of Human Genetics*, *81*, 939.

Chechik, G., Globerson, A., Tishby, N., Anderson, M.J., Young, E.D., Nelken, I. (2001). In T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), *Neaural information processing systems 14* (Vol. 1, p. 173) MIT Press.

Cover, T.M., & Thomas, J.A. (2006). Elements of information theory, 2nd edn. Wiley-Interscience.

DeWeese, M.R., & Meister, M. (1999). *Network: Computation in Neural Systems*, *10*, 325.

Fairhall, A., Shea-Brown, E., Barreiro, A. (2012). *Current Opinion in Neurobiology*, *22*, 653.

Flecker, B., Alford, W., Beggs, J.M., Williams, P.L., Beer, R.D. (2011). *Chaos*, *21*, 037104.

Fraser, A.M., & Swinney, H.L. (1986). *Phys. Rev. A*, *33*, 1134.

Fujisawa, S., Amarasingham, A., Harrison, M.T., G. Buzsáki (2008). *Nature Neuroscience*, *11*, 823.

Garofalo, M., Nieus, T., Massobrio, P., Martinoia, S. (2009). *PLoS One*, *4*, e6482.

Gat, I., & Tishby, N. (1999). In M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), *Neural information processing systems 11* (p. 111). MIT Press.

Globerson, A., Stark, E., Vaadia, E., Tishby, N. (2009). *PNAS*, *106*, 3490.

Gollisch, T., & Meister, M. (2008). *Science*, *319*, 1108.

Griffith, V., & Koch, C. (2012). *Quantifying synergistic mutual information*. arXiv:12054265v2.

Han, T.S. (1975). *Information and Control*, *29*, 337.

Han, T.S. (1978). *Information and Control*, *36*, 133.

Hatsopoulos, N., Geman, S., Amarasingham, A., Bienenstock, E. (2003). *Neurocomputing*, *52*, 25.

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J. (2007). *Physics Reports*, *441*, 1.

Honey, C.J., Kotter, R., Breakspear, M., Sporns, O. (2007). *PNAS*, *104*, 10240.

Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D., Yuste, R. (2004). *Science*, *304*, 559.

Ito, S., Hansen, M.E., Heiland, R., Lumsdaine, A., Litke, A.M., Beggs, J.M. (2011). *PLoS One*, *6*(21), e27431.

Jakulin, A., & Bratko, I. (2008). Quantifying and visualizing attribute interactions. arXiv:cs/0308002v3.

James, R.G., Ellison, C.J., Crutchfield, J.P. (2011). *Chaos*, *21*, 037109.

Kamioka, H., Maeda, E., Jimbo, Y., Robinson, H.P.C., Kawana, A. (1996). *Neuroscience Letters*, *206*, 109.

Kennel, M.B., Shlens, J., Abarbanel, H.D.I., Chichilnisky, E.J. (2005). *Neural Computation*, *17*, 1531.

Latham, P.E., & Nirenberg, S. (2005). *Journal of Neuroscience*, *25*, 5195.

Lizier, J.T., Heinzle, J., Horstmann, A., Haynes, J.D., Prokopenko, M. (2011). *Journal of Computational Neuroscience*, *30*, 85.

Lizier, J.T., Flecker, B., Williams, P.L. (2013). *Towards a synergy-based approach to measuring information modification*. arXiv:1303.3440.

Louie, K., & Wilson, M.A. (2001). *Neuron*, *29*, 145.

Lungarella, M., & Sporn, O. (2006). *PLoS One*, *2*, e144.

Madhavan, R., Chao, Z.C., Potter, S.M. (2007). *Physical Biology*, *4*, 181.

Marschinski, R., & Kantz, H. (2002). *European Physical Journal B*, *30*, 275.
Martignon, L., Deco, G., Laskey, K., Diamond, M., Freiwald, W., Vaadia, E. (2000). *Neural Computation*, *12*, 2621.
Matsuda, H. (2000). *Physical Review E*, *62*, 3096.
McGill, W.J. (1954). *Psychometrika*, *19*, 97.
Nemenman, I., Bialek, W., de Ruyter van Steveninck, R.R. (2004). *Physical Review E*, *69*, 056111.
Nirenberg, S., Carcieri, S.M., Jacobs, A.L., Latham, P.E. (2001). *Nature*, *411*, 698.
Ohiorhenuan, I.E., & Victor, J.D. (2011). *Journal of Computational Neuroscience*, *30*, 125.
Ohiorhenuan, I.E., Mechlar, F., Purpura, K.P., Schmid, A.M., Hiu, Q., Victor, J.D. (2010). *Nature Letters*, *466*, 617.
Olbrich, E., Bertschinger, N., Ay, N., Jost, J. (2008). *European Physical Journal B*, *63*, 407.
Optican, L.M., & Richmond, B.J. (1987). *Journal of Neurophysiology*, *57*, 162.
Paiva, A.R.C., Park, I., Principe, J.C. (2010). *Neural Computation and Application*, *19*, 405.
Paninski, L. (2003). *Neural Computation*, *15*, 1191.
Panzeri, S., & Treves, A. (1996). *Network: Computation in Neural Systems*, *7*, 87.
Panzeri, S., Petersen, R.S., Schultz, S.R., Lebedev, M., Diamond, M.E. (2001). *Neuron*, *29*, 769.
Panzeri, S., Senatore, R., Montemurro, M.A., Petersen, R.S. (2007). *Journal of Neurophysiology*, *98*, 1064.
Pasquale, V., Massobrio, P., Bologna, L.L., Chiappalonea, M., Martinoia, S. (2008). *Neuroscience*, *153*, 1354.
Pazienti, A., Maldonado, P.E., Diesmann, M., Grun, S. (2008). *Brain Research*, *1225*, 39.
Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., Simoncelli, E.P. (2008). *Nature*, *454*, 995.
Quiroga, R.Q., & Panzeri, S. (2009). *Nature Reviews Neuroscience*, *10*, 173.
Quiroga R.Q., & Panzeri S. (Eds.) (2013). *Principles of Neural Coding*. CRC Press LLC.
Rieke, F., Warland, D., de Ruyter van Steveninck, R.R., Bialek, W. (1997). *Spikes: exploring the neural code.* MIT Press.
Rivlin-Etzion, M., Ritov, Y., Heimer, G., Bergman, H., Bar-Gad, I. (2006). *Journal of Neurophysiology*, *95*, 3245.
Rokem, A., Watzl, S., Gollisch, T., Stemmler, M., Herz, A.V.M., Samengo, I. (2006). *Journal of Neurophysiology*, *95*, 2541.
Rolston, J.D., Wagenaar, D.A., Potter, S.M. (2007). *Neuroscience*, *148*, 294.
Schreiber, T. (2000). *Physical Review Letters*, *85*, 461.
Schneidman, E., Bialek, W., Berry II, M.J. (2003a). *Journal of Neuroscience*, *23*, 11539.
Schneidman, E., Still, S., Berry II, M.J., Bialek, W. (2003b). *Physical Review Letters*, *91*, 238701.
Schneidman, E., Berry II, M.J., Segev, R., Bialek, W. (2006). *Nature*, *440*, 1007.
Shannon, C.E. (1948). *The Bell System Technical Journal*, *27*, 379.
Shimazaki, H., Amari, S., Brown, E.N., Grun, S. (2012). *PLoS Computational Biology*, *8*(3), e1002385.
Shlens, J., Field, G.D., Gauthier, J.L., Grivich, M.I., Petrusca, D., Sher, A., Litke, A.M., Chichilnisky, E.J. (2006). *Journal of Neuroscience*, *26*, 8254.
Shlens, J., Kennel, M.B., Abarbanel, H.D.I., Chichilnisky, E.J. (2007). *Neural Computation*, *19*, 1683.
Sporns, O., Tononi, G., Edelman, G.E. (2000). *Cerebral Cortex*, *10*, 127.
Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., Bialek, W. (1997). *Physical Review Letters*, *80*, 197.
Tang, A., Jackson, D., Hobss, J., Chen, W., Smith, J.L., Patel, H., Prieto, A., Petrusca, D., Grivich, M.I., Sher, A., Hottowy, P., Dabrowski, W., Litke, A.M., Beggs, J.M. (2008). *Journal of Neuroscience*, *28*, 505.
Tetzlaff, C., Okujeni, S., Egert, U., Worgotter, F., Butz, M. (2010). *PLoS Computational Biology*, *6*, e1001013.
Timme, N., Alford, W., Flecker, B., Beggs, J.M. (2011). Multivariate information measures: an experimentalist's perspective. arXiv:1111.6857.
Tononi, G., Sporns, O., Edelman, G.M. (1994). *Proceedings of the National Academy of Sciences*, *91*, 5033.
Treves, A., & Panzeri, S. (1995). *Neural Computation*, *7*, 399.
Varadan, V., Miller III, D.M., Anastassiou, D. (2006). *Bioinformatics*, *22*, e497.
Vicente, R., Wibral, M., Lindner, M., Pipa, G. (2011). *Journal of Computational Neuroscience*, *30*, 45.
Victor, J.D. (2002). *Physical Review E*, *66*, 051902.
Victor, J.D. (2006). *Biological Theory*, *1*, 302.
Wagenaar, D.A., Pine, J., Potter, S.M. (2006a). *BMC Neuroscience*, *7*.
Wagenaar, D.A., Nadasdy, Z., Potter, S.M. (2006b). *Physical Review E*, *73*, 051907.
Wang, L., Narayan, R., Graña, G., Shamir, M., Sen, K. (2007). *Journal of Neuroscience*, *27*(3), 582.
Warland, D.K., Reinagel, P., Meister, M. (1997). *Journal of Neurophysiology*, *78*, 2336.
Watanabe, S. (1960). *IBM Journal of Research and Development*, *4*, 66.
Wennekers, T., & Ay, N. (2003). *Theory in Bioscience*, *122*, 5.
Williams, P.L., & Beer, R.D. (2010). Decomposing multivariate information. arXiv:1004.2515v1.
Williams, P.L., & Beer, R.D. (2011). Generalized measures of information transfer. arXiv:1102.1507v1.
Yeh, F.C., Tang, A., Hobbs, J.P., Hottowy, P., Dabrowski, W., Sher, A., Litke, A., Beggs, J.M. (2010). *Entropy*, *12*, 89.
Yu, S., Yang, H., Nakahara, H., Santos, G.S., Nikolic, D., Plenz, D. (2011). Higher-order interactions characterized in cortical activity. *Journal of Neuroscience*, *31*, 17514.
Ziv, J., & Lempel, A. (1977). *IEEE Transactions on Information Theory*, *23*, 337.