 doi:10.20944/preprints202002.0066.v2

Article

A Partial Information Decomposition Based on Causal Tensors

David Sigtermans [1]*

[1] ASML
* Correspondence: david.sigtermans@asml.com

Abstract: We propose a partial information decomposition based on the newly introduced framework of causal tensors, i.e., multilinear stochastic maps that transform source data into destination data. The innovation that causal tensors introduce is that the framework allows for an exact expression of an indirect association in terms of the constituting, direct associations. This is not possible when expressing associations only in measures like mutual information or transfer entropy. Instead of a priori expressing associations in terms of mutual information or transfer entropy, the a posteriori expression of associations in these terms results in an intuitive definition of a nonnegative and left monotonic redundancy, which also meets the identity property. Our proposed redundancy satisfies the three axioms introduced by Williams and Beer. Symmetry and self-redundancy axioms follow directly from our definition. The data processing inequality ensures that the monotonicity axiom is satisfied. Because causal tensors can describe both mutual information as transfer entropy, the partial information decomposition applies to both measures. Results show that the decomposition closely resembles the decomposition of other another approach that expresses associations in terms of mutual information a posteriori. A negative synergistic term could indicate that there is an unobserved common cause.

Keywords: information theory; causal inference; causal tensors; transfer entropy; partial information decomposition; left monotonicity; identity property; unobserved common cause

1. Introduction

The ability to decompose information within a multivariate system, i.e., systems comprising over two random variables, allows us to *diagnose* behavior of these systems. Mutual information, the information measure used in information theory [1], does not lead to a satisfactory decomposition. For example, the widely used "Interaction Information" [2] can have negative values, which is counterintuitive. An alternative approach is offered via "partial information decomposition" [3]. *Total information* is written as the sum of nonnegative information components. The definition of these nonnegative information components is still an open problem however, as summarized in [4].

In this article we contribute to this discussion by showing that a partial information decomposition in nonnegative contributors follows naturally from the framework of causal tensors, i.e., the set of transition probability matrices. For mutual information, the tensor can be represented by a matrix. The tensor comprises several transition probability matrices in the case of transfer entropy. Information theory [1] models the association between data as transmission of source data towards a destination via a *communication channel* towards. A channel is characterized by its probability transition matrix [5]. The association between the source data and the destination data is the mutual information (MI). If data is transmitted from a source to a mediator and from there to the destination, i.e., via a transmission path comprising three or more nodes, the mutual information between the source and destination can not be expressed in terms of the mutual informations between the source and mediator and the MI between the mediator and destination. Using the causal tensors however, the causal tensor of the resulting communication channel between source and destination is a function of the constituting communication channels along the transmission path.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

Instead of a priori expressing associations in terms of mutual information or transfer entropy, we propose to use a posteriori expression of associations in these terms. In doing so, an intuitive definition of redundancy and related, unique information arises.

2. Materials and Methods

Because causal tensors as a representation for the communication channel plays a central role, a short summary is given

2.1. Causal tensors

In information theory, the data are realizations of random variables representing stationary ergodic processes [1]. Because the data comprise elements from a finite alphabet, we can describe an outcome using the value, i.e., alphabet symbol, or using the index, assuming a fixed, e.g., lexicographic, order of the alphabet elements. In this article we use the latter.

The communication channel transforms the probability mass function (pmf) of the source data in the pmf of the destination data via a linear mapping. With p^j representing the j^{th} element of the destination pmf, and p^i representing the i^{th} element of the source pmf, the relation between source and destination is given by

$$p^j = \sum_i p^i A_i^j.$$

The elements of the tensor $\mathcal{A}$, i.e., A_i^j, represent the transition probabilities $A_i^j = p(j^{th} \ destination \ symbol | i^{th} \ source \ symbol)$. A communication channel is the conceptual implementation of the Law of Total Probability [6]. This immediately implies that the source pmf can be reconstructed from the destination pmf: $p^i = \sum_j p^j A_j^{\ddagger i}$. The ‡ indicates that the source pmf is reconstructed. This is relevant in case a source is distinguishable from a destination, i.e., when transfer entropy is used. The mutual between the source data, generated by process X, and the destination data, generated by the process Y, is expressed as

$$I(X,Y) = \sum_{i,j} p^{ij} \log_2 \left[\frac{A_i^j}{p^j} \right]. \tag{1}$$

Equation 1 is equivalent to standard expression for mutual information, $I(X,Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log_2 \left[p(y|x)/p(y) \right]$ (see, for example, [5]). This can be seen by switching from the index notation to the notation in alphabet elements by replacing the transition probability matrix elements A_i^j with the conditional probability $p(y|x)$, the joint probability p^{ij} with $p(x,y)$, and the output probability p^j with $p(y)$. The variable x is selected from alphabet $\mathcal{X}$, and the variable y is selected from the alphabet $\mathcal{Y}$ respectively.

Transfer entropy [7] is an information theoretical implementation of "Wieners principle of causality" [8]: a "cause" combined with the past of the "effect" predicts the effect better than that the "effect" predicts itself. It was proven that with a slight modification of the original proposed transfer entropy (TE) fully complies with Wieners principle of causality [9]. Transfer entropy is the measure of association for data transmission via a network of communication channels with an *inverse multiplexer* topology (see Figure 1a). This becomes apparent when transfer entropy is written as

$$TE_{X \to Y} = \sum_{\psi_g^- \in \mathcal{Y}^\ell} p(\psi_g^-) \sum_{\substack{\mathbf{x}^- \in \mathcal{X}^m \\ y \in \mathcal{Y}}} p(\mathbf{x}^-, y | \mathbf{y}^- = \psi_g^-) \log_2 \left[\frac{p(y|\mathbf{x}^-, \mathbf{y}^- = \psi_g^-))}{p(y|\mathbf{y}^- = \psi_g^-)} \right],$$

with $\mathbf{x}^-$ representing the cause selecting symbols from the alphabet $\mathcal{X}^m$, $\mathbf{y}^-$ representing the past of the effect, ψ_g^- equals the g^{th} alphabet element from the alphabet $\mathcal{Y}^\ell$, and y representing the effect.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

It is assumed that Y is a Markov process of order ℓ, and that the "cause" consists of m subsequent symbols.

Transition probability matrices can be associated with this inverse multiplexer channel. There are $|\mathcal{Y}^{\ell}|$ sub-channels, with $|\cdot|$ is used to indicate the cardinality of the alphabet. Per sub-channel, the communication channel is identical to a communication channel giving rise to MI as the measure of association, therefore, the elements of the g^{th} sub-channel of the causal tensor $\mathcal{A}$ are given by A^j_{gi}. Like MI, transfer entropy can be expressed in terms of the tensor:

$$TE_{X\to Y} = \sum_{g,i,j} p^{ijg} \log_2 \left[\frac{A^j_{gi}}{p^j_g} \right].$$

Mutual information also results from transmission over an inverse multiplexer, an inverse multiplexer comprising one sub-channel.

Definition 1. *Mutual information is the measure of association from data transmission via a single-channel inverse multiplexer. Transfer entropy is the measure of association from data transmission via a multi-channel inverse multiplexer.*

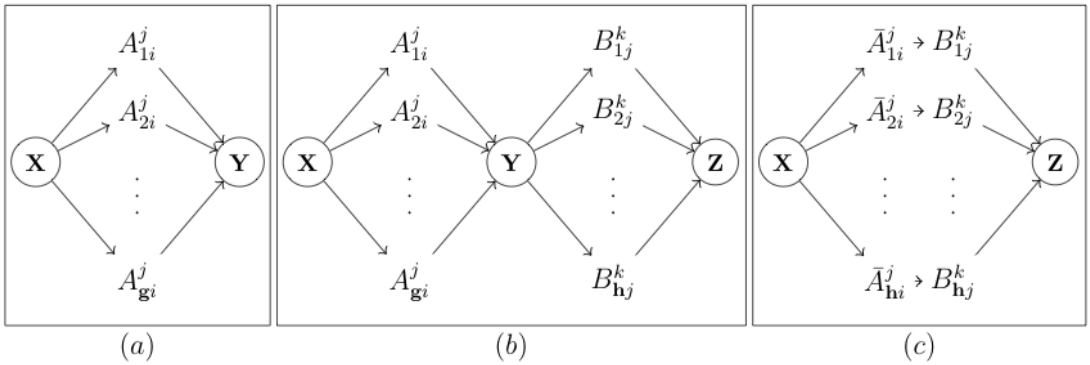

Figure 1. (**a**) The inverse multiplexer representing the communication network between X and Y. Source data is partitioned on the past of the effect, indicated by index g, and transmitted via the related communication channel. (**b**) The inverse multiplexers representing the transmission path $\{x, y, z\}$. Index h is related to the past of the z. (**c**) An equivalent representation network communication channels representing the transmission path $\{x, y, z\}$. The causal tensor elements $\bar{A}^j_{hi}$ equals $\sum_g p^g_{hi} A^j_{gi}$. Per "sub-channel", the resulting causal tensor equals $\sum_j \bar{A}^j_{hi} B^k_{hj}$.

2.1.1. Causal Tensors of a Cascade

A system comprising more than one random variables can be expressed as a graph in which the nodes represent the random variables, and the edges represent the existence of an association. The edges are undirected when MI is used as a measure of association and directed when TE is used. This gives rise to "transmission paths".

Definition 2. *A transmission path, or path in short, is defined as the sequence in which nodes have been used to transmit the data. The source is defined as the first variable in the path, the destination is the last variable in the path. A path is denoted as* $\{source, mediator_1, \cdots, mediator_n, destination\}$.

The association between a source and destination of a path comprising over two nodes, i.e., a cascade of (direct) paths, cannot be expressed in terms of the association constituting direct paths when MI or TE are used. However, the following theorem can be proven [10]:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

Theorem 1. *The causal tensor of a cascade of paths can be expressed in terms of the causal tensors of the constituting direct paths.*

For example, assume that the tensor elements for the path $\{x, y\}$ are given by A^j_{gi}, with g the index for the past of the effect (y), and the tensor elements for the path $\{x, y\}$ are given by B^k_{hj}, with h the index for the past of the effect (z). The causal tensor $\mathcal{T}$ for the path $\{x, y, z\}$ equals

$$\mathcal{T}\{x, y, z\} = \sum_{g,j} p^g_{hi} A^j_{gi} B^k_{hj}. \tag{2}$$

Figure 1b depicts the transmission of data over two inverse multiplexers in series, resulting in Eq.(2). The term $\sum_g p^g_{hi} A^j_{gi}$ can be interpreted as the weighted sum of the causal tensors of the sub-channels of the first direct path, $\{x, y\}$, and it evaluates to a causal tensor $\bar{A}^j_{hi}$ [10]. Transmission of data over a cascade of multi-channel inverse multiplexers is equivalent to the transmission of data over a parallel set of cascades of single-channel inverse multiplexers, i.e., results applicable to MI also apply to TE. For this reason, we can restrict ourselves, without loss of generality, to MI. The mutual information for the path $\{x, y, z\}$, the tensor elements of the resulting causal tensor equal $\sum_j A^j_i B^k_j$,

$$I\{x, y, z\} = \sum_{i,j,k} p^{ijk} \log_2 \left[\frac{\sum_j A^j_i B^k_j}{p^k} \right].$$

2.1.2. Indirect Associations and No Associations

There are two underlying reasons to delete an edge in a graph: *(i)* the association it represents is indirect, or *(ii)* there is no association at all. These two reasons can be distinguished using causal tensors. First, because the causal tensor of a cascade can be determined exactly from the constituting causal tensors, causal tensors can differentiate between direct and indirect associations [10].

Theorem 2. *If the association between two nodes is indirect, the causal tensor of the direct path equals the resulting causal tensor of the cascade. The direct path does not exist.*

For example, the path $\{x, z\}$ does not exist in when the graph $X \to Y \to Z$ is the ground truth. Second, if there is no association between two nodes at all, the causal tensor represents a communication channel that cannot transmit any information. Here the transition probability matrix has identical rows, e.g., $\forall\, i \neq f :\ A^j_i = A^j_f$. In this case the direct path exists neither.

2.2. Partial Information Decomposition

The partial information decomposition framework of Williams and Beer [3] allows for a decomposition of the total information in nonnegative unique, redundant, and synergistic information components. The unique information $\mathcal{U}(Y; Z)$ represents information in Z only provided by Y and not by X. The redundant information $\mathcal{R}(X, Y; Z)$ represents the information in Z provided by both X and Y. The synergistic information $\mathcal{S}(X, Y; Z)$ represents information in Z that results via interaction between X and Y. The relations between these information components for a system comprising three variables are given by the following set of equations,

$$I(X, Y; Z) = \mathcal{U}(Y; Z) + \mathcal{U}(X; Z) + \mathcal{R}(X, Y; Z) + \mathcal{S}(X, Y; Z), \tag{3}$$
$$I(Y; Z) = \mathcal{U}(Y; Z) + \mathcal{R}(X, Y; Z), \tag{4}$$
$$I(X; Z) = \mathcal{U}(X; Z) + \mathcal{R}(X, Y; Z). \tag{5}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

Williams and Beer propose three redundancy related axioms: *(i)* Symmetry: redundancy does not change when sources are permuted, e.g., $R(Z;X,Y) = R(Z;Y,X)$. *(ii)* Self-Redundancy: for a single source, the redundancy equals the mutual information between the source and the destination, e.g., $R(Z;X) = I(Z;X)$. *(iii)* Monotonicity: the redundancy does not increase when a new source is added, e.g., $R(Z;X) \geq R(Z;X,Y)$.

2.2.1. Redundancy, Indirect Paths and the Data Processing Inequality

If the association between two nodes is indirect, no direct path exists. Source data is transmitted to the mediator node, which stores, possibly copies, possibly modifies, and possibly enriches the received information with information from this mediator node, after which it is transmitted again towards the next mediator node or destination node. This consideration leads to the following proposition:

Proposition 1. *Unique information can only result from data transmission via a direct path. Redundant information is the consequence of data transmission via an indirect path.*

A direct consequence of this proposition is that in case the chain $X \to Y \to Z$ is the ground truth, no unique information is shared between X and Z, or stated otherwise, all information shared between X and Z is redundant. Another immediate consequence of this proposition is that in case of the XOR example from [11], there is neither unique, nor redundant information shared between the sources and the destination: all information shared is synergistic.

In a fully connected three-node system, there are two source nodes transmitting data to the destination node, and per source node, there is one indirect path between that source node and the destination node. All indirect paths fall in two categories: *(i)* the indirect path includes the direct path between a specific source and destination, and *(ii)* the indirect path does *not* include the direct path between a specific source and destination. For the first category, the Data Processing Inequality (PID) [5] is directly applicable. Data Processing Inequality states that processing of data can never increase the amount of information. For the path $\{x,y,z\}$ this means that $I\{x,y,z\} \leq \min[I\{x,y\}, I\{y,z\}]$. Via this path the redundant information from X is transmitted. For the path $\{y,x,z\}$ this means that $I\{y,x,z\} \leq \min[I\{y,x\}, I\{x,z\}]$. Via this path the redundant information from Y is transmitted. Based on this example, the following definition of redundancy is proposed:

Definition 3. *Redundant information shared between a specific set of sources with respect to a destination is defined as the weakest of all indirect paths that: (1) contain all the sources, (2) start with a source and ends at the destination, and (3) do not contain non-existing paths.*

Because the redundancy in this definition equals a mutual information, the proposed redundancy is per definition nonnegative. This definition of redundancy also satisfies the three axioms introduced in [3].

Sketch of Proof of Symmetry. Because of the definition of redundancy, all indirect paths representing all permutations of sources are compared. The order of the sources in the redundancy expression is therefore irrelevant. □

Sketch of Proof of Self-Redundancy. Assume that one source is a copy of the other source, e.g., $Y = X$. This means that for the causal tensors describing the mapping between the sources and destination, $\mathcal{B}$ and $\mathcal{C}$: $\mathcal{B} = \mathcal{C}$. Per definition $R(X,X;Z) = I\{x,x,z\}$. The causal tensor for the path $I\{x,x\}$, $\mathcal{T}I\{x,x\}$, equals the Kronecker delta δ_i^j: $\delta_1^j = 0$ unless $i = j$ in which case $\delta_i^j = 1$. Therefore, the causal tensor for the path $I\{x,x,z\}$ equals $\mathcal{B}$, i.e., $I\{x,x,z\} = I(X;Z)$. □

Sketch of Proof of Monotonicity. Because of the definition of redundancy, adding more sources increases the number of nodes within the indirect path. As per Data Processing inequality, the adding sources can never increase the redundancy. □

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

Apart from these three axioms, Bertschinger et al. [12] and Harder et al. [13] proposed other properties. The proposed redundancy measure satisfies both the "left monotonicity" property and the "identity property".

2.2.2. Left Monotonicity Property

Left monotonicity captures the intuition that if X and Y share some information about Z_1, "then at least the same amount of information is available to reduce the uncertainty about the joint outcome" of Z_1 and Z_2 [12]:

$$R(Z_1; Y, X) \leq R(Z_1 Z_2; Y, X).$$

For readability, the proof of the left monotonic property has been moved to Appendix A.

2.2.3. Identity Property

The intuition behind the identity property [13] is that if the destination is a join of the inputs, the redundancy equals the mutual information of inputs, i.e.,

$$R(XY; X, Y) = I(Y; X),$$

with $XY = X \cup Y$. In Appendix B we prove that the proposed redundancy property satisfies the identity property.

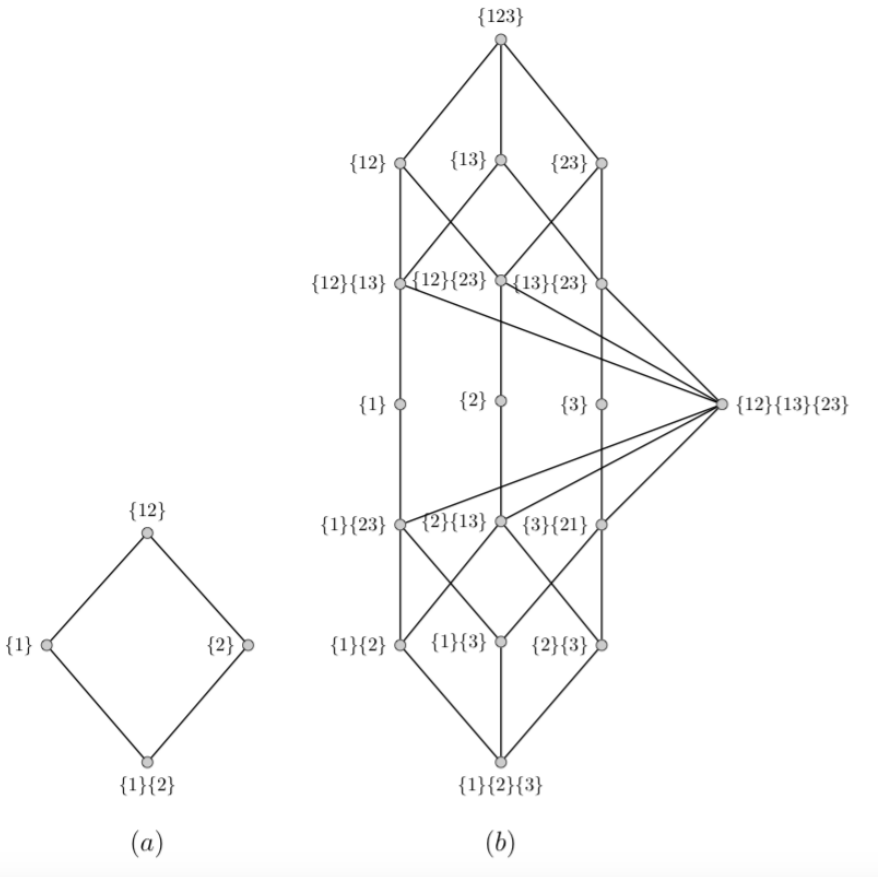

Figure 2. The redundancy lattice [3] for: **a**) two sources, and (**b**) three sources. In case two lattice nodes are connected, then the redundancy related to the highest lattice node (in position) is greater than or equal to the redundancy of the lower lattice node.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

For a three-node system, the proposed redundancy gives rise to an identical redundancy lattice introduced by [3], and shown in Figure 2. Because of the DPI there is a natural ordering in terms of (self-) redundancy:

- $R(X,X;Z) \geq R(X,Y;Z)$,
- $R(Y,Y;Z) \geq R(X,Y;Z)$,
- $R(\{X,Y\},\{X,Y\}) \geq R(Y,Y;Z)$,
- $R(\{X,Y\},\{X,Y\}) \geq R(X,X;Z)$.

The last two relations are a consequence of the fact that the tensor of the path $\{x,z\}$ can be expressed as a tensor contraction of the tensors related to the paths $\{xy,z\}$ and $\{x,y\}$, and that the tensor of the path $\{y,z\}$ can be expressed as a tensor contraction involving the tensors related to the path $\{xy,z\}$ and $\{y,x\}$ [10].

2.2.4. Unique Information

Unique information for a three-node system is defined by Eq.(4) and Eq.(5). Using the proposed redundancy, the resulting unique information is nonnegative because of the DPI. The proof is straightforward and left to the reader. The unique information defined in this fashion, fully matches our intuition. This can demonstrated by applying it to the fully connected three-node system. The information in Z results from information transmitted via the paths $I\{x,y,z\}$ and $I\{y,x,z\}$. The redundant information $\mathcal{R}(X,Y;Z)$ equals

$$\mathcal{R}(X,Y;Z) = \min\left[I\{x,y,z\}, I\{y,x,z\}\right].$$

The causal tensor elements of the paths $I\{y,x,z\}$ and $I\{x,y,z\}$ are given by $\sum_i A_j^{\ddagger i} C_i^k$ and $\sum_j A_i^j B_j^k$ respectively. Using Eq.(4) the unique information we get

$$\mathcal{U}(Y;Z) = \max\left[\sum_{i,j,k} p^{ijk} \log_2\left[\frac{B_j^k}{\sum_m A_j^{\ddagger m} C_m^k}\right], \sum_{i,j,k} p^{ijk} \log_2\left[\frac{B_j^k}{\sum_\ell A_i^\ell B_\ell^k}\right]\right]. \tag{6}$$

The first sum is a measure for the divergence between the direct path $\{y,z\}$ and the indirect path $\{y,x,z\}$. If the *association* between Y and Z is indirect, this sum evaluates to zero. This also implies that a direct path between Y and Z does not exist. The second sum is an indication how much Y differs from an exact copy of X. If Y is an exact copy, this term evaluates to zero: no unique information is shared between Y and Z because all information was already shared via X.

2.2.5. Relationship of Proposed Redundancy with Redundancy Lattice

As shown earlier, the proposed redundancy for two sources matches the redundancy lattice derived in [3]. With three sources, this is also the case. The lattice nodes are indicated as $\{X_1\}\{\cdots\}\{X_n\}$ (see Figure 2). The redundancy related to the lattice node equals $R(X_1,\cdots,X_n;Z)$, i.e., the minimum overall mutual informations of the indirect paths ending in Z and containing all permutations of the sources $X_1,\cdots,X_n$. This means that a lattice node gives a description of all individual sources involved. Because of the DPI, the redundancy associated with a lattice node must be less than or equal to the redundancy of lattice nodes comprising only a subset of the sources. For example, $R(1,23;Z) \leq R(1;Z)$, where 23 indicates the join of the sources 2 and 3. When this is combined with the order suggested by the redundancy lattice for two sources, e.g., $R(12;Z) \geq R(1;Z)$, the proposed definition fully complies with the order implied by the redundancy lattice.

3. Results

In this section the behavior of the proposed partial information decomposition we start with investigating its behavior with respect the conceptual issue related to, I_{min}, the original redundancy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

measure used in [3]: I_{min} does not distinguish between "same information" or "the same amount of information".

3.1. Two Bit Copy Problem

A conceptual problem with the redundancy measure used in [3], I_{min}, is illustrated with the so called "two-bit copy problem". For two independent and identically distributed binary variables X and Y, the destination Z is a copy of these two variables: $Z = (X, Y)$. It can be shown that $I_{min}(Z; \{1\}\{2\}) =$ 1 bit [14]. The problem lies in the fact that there is no overlap between the information of both variables: the result does not match out intuition, I_{min} seems to overestimate the redundancy. All indirect paths

Table 1. (Distribution for the "two-bit copy problem".

X	Y	Z	$p(Z)$
0	0	(0,0)	$\frac{1}{4}$
0	1	(0,1)	$\frac{1}{4}$
1	0	(1,0)	$\frac{1}{4}$
1	1	(1,1)	$\frac{1}{4}$

contain the paths between X and Y. Using Table 1, the causal tensor for the path $\{x, y\}$, $\mathcal{A}$, can be determined:

$$\mathcal{A} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The tensor related to the path $\{y, x\}$, $\mathcal{A}^{\ddagger}$ equals $\mathcal{A}$. This implies that no information can be transmitted via the paths between X and Y, i.e., $R(X, Y; Z) = 0$. The proposed redundancy does not suffer from this conceptual issue, because the proposed redundancy satisfies the identity property.

3.2. Dyadic and Triadic Systems

Next we apply the proposed method to the two data sets from Table 2. Although these sets have different underlying dependency structures, they apparently have the same statistical structure [15].

Table 2. Two systems, both comprising three random variables with identical joint probabilities per combination of the random variables. The underlying structures are very different, which can be seen when the variables are represented in two bits, e.g., the binary expansion for $X = 3$ equals $X_0X_1 = 11$. **(a)** For the dyadic (pair-wise) set, $X_0 = Y_1, Y_0 = Z_1$, and $Z_0 = X_1$. **(b)** For the triadic (three-way) set, $X_0 + Y_0 + Z_0$ mod2, and $X_1 + Y_1 + Z_1$.

(a) Dyadic				**(b) Triadic**			
X	Y	Z	p	X	Y	Z	p
0	0	0	$\frac{1}{8}$	0	0	0	$\frac{1}{8}$
0	2	1	$\frac{1}{8}$	1	1	1	$\frac{1}{8}$
1	0	2	$\frac{1}{8}$	0	2	2	$\frac{1}{8}$
1	2	3	$\frac{1}{8}$	1	3	3	$\frac{1}{8}$
2	1	0	$\frac{1}{8}$	2	0	2	$\frac{1}{8}$
2	3	1	$\frac{1}{8}$	3	1	3	$\frac{1}{8}$
3	1	2	$\frac{1}{8}$	2	2	0	$\frac{1}{8}$
3	3	3	$\frac{1}{8}$	3	3	1	$\frac{1}{8}$

For the dyadic set, the causal tensors are given by:

$$\mathcal{A} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}, \mathcal{C} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

$\mathcal{B}=\mathcal{A}$, $\mathcal{A}^{\ddagger}=\mathcal{C}$, $\mathcal{B}^{\ddagger}=\mathcal{C}$, and $\mathcal{C}^{\ddagger}=\mathcal{A}$. Because no relation is the result of a cascade, e.g., $B_j^k \neq \sum_i A_j^{\ddagger i} C_i^k$, the structure is that of an undirected triangle. Lets assume we are interested in the partial information decomposition of the total information in Z. Because $\forall\ i,k:\ \sum_j A_i^j B_j^k = \frac{1}{4}$, the redundant information equals zero. For the triadic set, the causal tensors are given by:

$$\mathcal{A} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix},$$

$\mathcal{B}=\mathcal{A}$, $\mathcal{C}=\mathcal{A}$, $\mathcal{A}^{\ddagger}=\mathcal{A}$, $\mathcal{B}^{\ddagger}=\mathcal{A}$, and $\mathcal{C}^{\ddagger}=\mathcal{A}$. Here, the ground structure is that of a chain because any relation results from a cascade, e.g., $C_i^k = \sum_j A_i^j B_j^k$. To compare the partial information decomposition of the triadic set, we have to investigate the set of the chains $X \to Y \to Z$ and $Y \to X \to Z$. By definition, Z does not contain unique information from X, nor does it contain any unique information from Y.

This example shows that the difference in underlying structure is reflected in two ways. First, the graphs related to the dyadic set and the triadic set are different (a triangle versus a chain). Second, for the dyadic set there is no redundant information and no synergistic information, because the total information in Z can not exceed 2 bit. In the triadic set, Z contains only redundant and synergistic information.

3.3. Comparison with Other Measures

To get an idea about the behavior of the proposed redundancy measure compared three other measures: *(i)* the earlier mentioned redundancy measure I_{min}, *(ii)* I_{broja}, the redundancy measure proposed in [16], and *(iii)* the redundancy based on Pointwise Common Change in Surprisal, I_{CCS}[14]. For a description of the examples we refer to [14]. The proposed redundancy measure is represented as $I_{\triangle}$.

Table 3. PID for 5A.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
{12}	0.3333	0	0.1383	0.1383
{2}	0.3333	0.6666	0.5283	0.5283
{1}	0.3333	0.6666	0.5283	0.5283
{1}{2}	0.5850	0.2516	0.3900	0.3900

Table 4. PID for 5B.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
{12}	0.5	0	0	0
{2}	0.5	1	1	1
{1}	0	0.5	0.5	0.5
{1}{2}	0.5	0	0	0

Table 5. PID for 5C.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
{12}	0.67	0.67	0.67	0.67
{2}	0.25	0.25	0.25	0.25
{1}	0	0	0	0
{1}{2}	0	0	0	0

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

Table 6. PID for REDUCEDOR.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
$\{12\}$	0.69	0.69	0.38	0.40
$\{2\}$	0	0	0.31	0.29
$\{1\}$	0	0	0.31	0.29
$\{1\}\{2\}$	0.31	0.31	0	0.02

Table 7. PID for XOR.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
$\{12\}$	1	1	1	1
$\{2\}$	0	0	0	0
$\{1\}$	0	0	0	0
$\{1\}\{2\}$	0	0	0	0

Table 8. PID for AND/OR.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
$\{12\}$	0.5	0.5	0.29	0.19
$\{2\}$	0	0	0.21	0.31
$\{1\}$	0	0	0.21	0.31
$\{1\}\{2\}$	0.31	0.31	0.10	0

Table 9. PID for SUM.

Lattice Node	$I_\partial(I_{min})$	$I_\partial(I_{broja})$	$I_\partial(I_{CCS})$	$I_\partial(I_\triangle)$
$\{12\}$	1	1	0.5	0.5
$\{2\}$	0	0	0.5	0.5
$\{1\}$	0	0	0.5	0.5
$\{1\}\{2\}$	0.5	0.5	0	0

From these examples, it is clear that our proposed PID closely resembles the PID proposed by Ince [14]. This should not come as a surprise because the pointwise approach suggested by Ince expresses associations in terms of mutual information a posteriori.

3.4. Negative Synergistic Contributions Due to Unobserved Common Causes?

Table 10. Example of a negative synergistic component in the case of an unobserved common cause. **(a)** Data set consisting of three parameters. **(b)** Hidden common cause. $X = \tilde{Z}_1$, $Y = \tilde{Z}_1 \text{OR} \tilde{Z}_2$, and $Z = \tilde{Z}_1 \text{AND} \tilde{Z}_2$.

(a) Data set				(b) Hidden common cause
X	Y	Z	p	$\tilde{Z} = \tilde{Z}_1\tilde{Z}_2$
0	0	0	$\frac{1}{4}$	00
0	1	0	$\frac{1}{4}$	01
1	1	0	$\frac{1}{4}$	10
1	1	1	$\frac{1}{4}$	11

When the partial information decomposition is performed for the distribution in Table 10(a), the resulting synergistic term is negative. The redundancy $R(Z; X, Y)$ equals 0.0271 bit. For this system the interaction information, $I(X;Y) - I(X;Y|Z)$ [2], equals 0.1226 bit, i.e., the synergistic term

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

is negative. At this moment it is unclear if there are other reasons for negative synergistic contributions.

In this article, we have shown that a partial information decomposition comprising nonnegative unique and redundant contributions follows naturally from the framework of causal tensors. Because we introduced no new information theoretical measures, it is our contention that a partial information decomposition is possible within the framework of "classical" Shannon information theory. The partial information decomposition is problematic when no exact expressions for indirect paths can be determined. It reduces to a rather straightforward exercise when this is possible, for example, within the framework of causal tensors. The potential non-negativity of synergistic information warrants future research. **Funding:** This research received no external funding.

Acknowledgments: I would like to thank Ryan James for asking relevant questions and providing me with challenging examples.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A.

Sketch of Proof of Left Monotonicity Property 2.2.2. Assume that there are two "destinations", Z_1 and Z_1. Furthermore, assume that the redundant information in destination Z_1 equals the mutual information of the path $\{x, y, z_1\}$:

$$R(Z_1; Y, X) = \sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j A_i^j B_j^{k_1}}{p^{k_1}} \right].$$

The index k_1 is related to destination Z_1. We multiplying the denominator and the numerator with the conditional probability $p_{k_1}^{k_2}$, the index k_2 is related to destination Z_2. With $p_{k_1}^{k_2} = \sum_{j'} p^{j'} p_{j'k_1}^{k_2}$, this results in

$$R(Z_1; Y, X) = \sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j \sum_{j'} A_i^j B_j^{k_1} p^{j'} p_{j'k_1}^{k_2}}{p^{k_1 k_2}} \right],$$

which is equivalent to

$$R(Z_1; Y, X) = \sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j A_i^j B_j^{k_1 k_2} p^j}{p^{k_1 k_2}} + \frac{\sum_j \sum_{j' \neq j} A_i^j B_j^{k_1} p^{j'} p_{j'k_1}^{k_2}}{p^{k_1 k_2}} \right]. \tag{A1}$$

Because the logarithm is a concave function, and the product of two probabilities is always less than or equal to the individual terms of the product, $p^{j'} p_{j'k_1}^{k_2} \leq p^{j'}$, Eq.(A1) leads to the following inequality:

$$R(Z_1; Y, X) \leq \sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j A_i^j B_j^{k_1 k_2} p^j}{p^{k_1 k_2}} + \frac{\sum_j \sum_{j' \neq j} A_i^j B_j^{k_1} p^{j'}}{p^{k_1 k_2}} \right]. \tag{A2}$$

According to the "pigeonhole principle", $\forall\, j' \neq j:\ p^{j'} \leq 1 - p^j$. Because the sum of products is less than or equal to the product of sums, the right-hand side of Eq.(A2) is less than, or equal to

$$\sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j \left(A_i^j B_j^{k_1 k_2} \right) \left(\sum_j p^j \right)}{p^{k_1 k_2}} + \frac{\left(\sum_j A_i^j B_j^{k_1} \right) \left(1 - \sum_j p^j \right)}{p^{k_1 k_2}} \right].$$

Because $\sum_j p^j = 1$, we finally get the inequality

$$R(Z_1; Y, X) \leq \sum_{i,k_1,k_2} p^{i,k_1,k_2} \log_2 \left[\frac{\sum_j A_i^j B_j^{k_1 k_2}}{p^{k_1 k_2}} \right].$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2020 doi:10.20944/preprints202002.0066.v2

The right-hand side equals the mutual information of the path $\{x, y, z_1 z_2\}$. In the same fashion it can be show that $R(Z_1; Y, X) \leq I\{y, x, z_1 z_2\}$, which finally proofs the left monotonicity of the proposed redundancy.

$$R(Z_1; Y, X) \leq R(Z_1 Z_2; Y, X).$$

□

Appendix B.

Sketch of Proof of Identity Property 2.2.3. Per definition

$$R(XY; X, Y) = \min\left[I\{x, y, xy\}, I\{y, x, xy\}\right].$$

Assume that, in index notation, p^i equals the pmf of X, p^j equals the pmf of Y, and p^{ij} equals the pmf of XY. Assume furthermore that A_i^j are the tensor elements for the path $\{x, y\}$. Per definition $A_j^{\ddagger i}$ represent the tensor elements for the path $\{y, x\}$.

Let tensor elements for the path $\{x, xy\}$ equal $C_{i'}^{ij}$, and the tensor elements for the path $\{y, xy\}$ equal $B_{j'}^{ij}$. Using these tensors, the redundancy equals

$$R(XY; X, Y) = \min\left[\sum_{i,j} p^{ij} \log_2\left[\frac{\sum_{j'} A_i^{j'} B_{j'}^{ij}}{p^{ij}}\right], \sum_{i,j} p^{ij} \log_2\left[\frac{\sum_{i'} A_j^{\ddagger i'} C_{i'}^{ij}}{p^{ij}}\right]\right].$$

The tensor elements $B_{j'}^{ij}$ are related to $A_j^{\ddagger i}$. It follows immediately that $B_{j'=j}^{ij} = A_j^{\ddagger i}$ and $B_{j'\neq j}^{ij} = 0$, Likewise, $C_{i'=i}^{ij} = A_i^j$ and $C_{i'\neq i}^{ij} = 0$. We can rewrite the joint probability as $p^{ij} = p^i A_i^j$, and as $p^{ij} = p^j A_j^{\ddagger i}$. With these expressions, we can rewrite the equation for the redundancy:

$$R(XY; X, Y) = \min\left[\sum_{i,j} p^{ij} \log_2\left[\frac{A_j^{\ddagger i}}{p^i}\right], \sum_{i,j} p^{ij} \log_2\left[\frac{A_i^j}{p^j}\right]\right].$$

Because both sums represent $I(X; Y)$, the mutual information between X and Y, the proposed redundancy measure satisfies the identity property. □

1. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x]. doi:10.1002/j.1538-7305.1948.tb01338.x.
2. McGill, W. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
3. Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. *preprint* **2010**, *1004*.
4. Lizier, J.; Bertschinger, N.; Wibral, M. Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work. *Entropy* **2018**, *20*, 307. doi:10.3390/e20040307.
5. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 1991.
6. Papoulis, A.; Pillai, S.U. *Probability, Random Variables, and Stochastic Processes*, fourth ed.; McGraw Hill, 2002.
7. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. doi:10.1103/PhysRevLett.85.461.
8. Beckenbach, E.F. *Modern mathematics for the engineer: second series*; New York : McGraw-Hill, 1961.
9. Wibral, M.; Pampu, N.; Priesemann, V.; Siebenhühner, F.; Seiwert, H.; Lindner, M.; Lizier, J.T.; Vicente, R. Measuring information-transfer delays. *PloS one* **2013**.
10. Sigtermans, D. Towards a Framework for Observational Causality from Time Series: When Shannon Meets Turing, 2020, [202001.0106]. doi:10.20944/preprints202001.0106.v1.

11. James, R.G.; Barnett, N.; Crutchfield, J.P. Information Flows? A Critique of Transfer Entropies. *Physical Review Letters* **2016**, *116*. doi:10.1103/physrevlett.116.238701.
12. Bertschinger, N.; Rauh, J.; Olbrich, E. Shared Information – New Insights and Problems in Decomposing Information in Complex Systems. *Springer Proceedings in Complexity* **2012**. doi:10.1007/978-3-319-00395-5_35.
13. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Physical review. E, Statistical, nonlinear, and soft matter physics* **2013**, *87*, 012130. doi:10.1103/PhysRevE.87.012130.
14. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **2016**, *19*. doi:10.3390/e19070318.
15. James, R.; Crutchfield, J. Multivariate Dependence Beyond Shannon Information. *Entropy* **2016**, *19*. doi:10.3390/e19100531.
16. Bertschinger, N.; Rauh, J.; Olbrich, E.; Ay, N. Quantifying Unique Information. *Entropy* **2013**, *16*. doi:10.3390/e16042161.