 doi:10.20944/preprints202001.0106.v1

Article

Towards a Framework for Observational Causality From Time Series: When Shannon Meets Turing

David Sigtermans [1]*

[1] ASML

* Correspondence: david.sigtermans@asml.com

Abstract: We propose a novel tensor-based formalism for inferring causal structures from time series. An information theoretical analysis of *transfer entropy (TE)*, shows that *TE* results from transmission of information over a set of communication channels. Tensors are the mathematical equivalents of these multi-channel causal channels. A multi-channel causal channel is a generalization of a *discrete memoryless channel (DMC)*. We consider a *DMC* as a single-channel causal channel. Investigation of a system comprising three variables shows that in our formalism, bivariate analysis suffices to differentiate between direct and indirect relations. For this to be true, we have to combine the output of multi-channel causal channels with the output of single-channel causal channels. We can understand this result when we consider the role of noise. Subsequent transmission of information over noisy channels can never result in less noisy transmission overall. This implies that a *Data Processing Inequality (DPI)* exists for transfer entropy.

Keywords: information theory; transfer entropy; time-delayed mutual information; data processing inequality; time series; causal tensor

1. Introduction

1.1. Motivation and Significance of the Work

Exact knowledge about the causal relationships that determine the behavior of complex systems is a holy grail in the (applied) sciences and engineering. This knowledge enables us to determine potential causes of certain effects, and it allows us to predict the effect of changes in causes (for example via simulation). In other words, it allows for causal inference and causal discovery [1]. In a causal relation, the cause precedes the effect (temporal precedence), and the cause physically influences the effect [2]. A causal description is essentially different from a description via statistical associations as illustrated by the adage "*correlation does not imply causation*". Examples of wrong, expensive, or even worse, harmful conclusions and policies based on statistical associations are part of common lore.

For a causal description, intervention is required [3]. These interventions enable us to differentiate between direct and indirect, or spurious, associations [1]. Because interventions are not always possible, we have to make do with observational data. A plethora of methods to infer causal structures from observational data have been developed, see for example [4–13]. What most these methods have in common is that they express relations via the causal effect, i.e., point-wise estimators that characterize "strength" of the association between a cause and an effect.

We propose a novel approach inspired by Turing machines [14]. If a human "computer" can decide, given the data, if a relation is causal, a Turing machine exists that reaches this decision in a

[1] An indirect association is an association via one or more mediators.

 doi:10.20944/preprints202001.0106.v1

mechanical way [15]. This is not a tautology, a Turing machine encodes the underlying principles leading to the decision that a relation is causal. It closely relates our approach to Structural Causal Models [16]. Instead of using point-wise estimators like transfer entropy [8] or time-delayed mutual information [5], we use stochastic tensors, i.e. multilinear maps.

Under the assumption that there are no hidden causes, unmeasured common causes or confounders, our formalism can differentiate between direct and indirect associations. We show that noise has a fundamental and, from the viewpoint of detecting spurious associations, a functional role. It is as if noise acts like "soft" interventions [17]. A surprising result is that for time-delayed mutual information *bivariate analysis suffices* to differentiating between direct and indirect associations. This result contradicts the long-held belief that this is impossible (see for example, [18] and [19]). The formalism furthermore allows for a simple proof of a Data Processing Inequality [20] for transfer entropy. This DPI can identify potential indirect relations when using TE, see for example [21].

1.2. Outline

The proposed formalism relies heavily on probability theory [22] and aspects and terminology used in causal inference [3]. In Section 1.3 we give a short overview of the most important ones.

To derive our formalism, we apply concepts from information theory [23] to transfer entropy. Transfer entropy is a measure that *can* capture causal relations, as far as encoded in the probability density functions, see for example [24,25]. In Section 2 we therefore introduction the applicable aspects of information theory, e.g., transmission of information, mutual information, communication channels and the tensor representation of communication channels. A tensor is a multilinear map that transforms an input into an output. We introduce transfer entropy in Section 3. We then show that transfer entropy is the average mutual information resulting from transmission of information of a set of communication channels. We call this set of channels *multi-channel causal channels*. Tensors are the mathematical equivalent to the set of channels. Using these tensors, we establish calculation rules in Section 4. We restrict ourselves to a system comprising three variables. We derive these results using index notation in Section 4. The result allow for a different notation that helps us to avoid a notational jungle of indices. This notation is borrowed from quantum mechanics and we introduce it in Section 5. This let us incorporate the temporal relations between a source and a destination, reflecting the additivity of interaction delays. Using this notation, we discuss some relevant findings for causal inference in Section 6.

In Section 7 we present two experiments that illustrate that our formalism can indeed detect nonlinear relationships and an underlying structure.

For readability, we have moved the longer proofs or sketches of proofs to the Appendix.

1.3. Preliminaries

Statistical independence is foundational to causal inference [7]. We will summarize the two most related and relevant assumptions: (1) the faithfulness assumption. (2) the Causal Markov Condition. A directed graph is said to be faithful to the underlying probability distributions if the independence relations that follow from the graph are the same independence relations that follow from the underlying probability distributions. For the chain $X \to Y \to Z$ the faithfulness assumption implies that X and Z are independent given Y. We denote this as $X \perp\!\!\!\perp Z|Y$.

The Causal Markov Condition states that a process is independent of its non-effects given its direct causes or its parents. This is relevant in the context of time series. A straightforward interpretation of this condition is that if the set of variables blocks all (undirected) paths between two variables, these two variables are independent given the set of variables blocking all paths [3]. We illustrate the Causal Markov Condition with an example that will be used later in this article.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

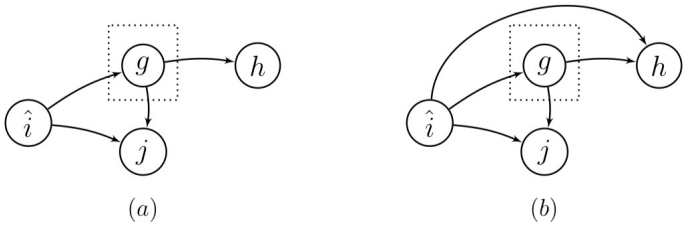

Figure 1. The graphs used in Example 1. (**a**) The dotted box blocks all paths between the non-parents $\hat{i}$ and j of h, and g is the only parent of h. (**b**) The dotted box does not block the path between $\hat{i}$ and h.

Example 1. *Lets start with the graph depicted in Figure 1(**a**). According to the Causal Markov Condition,* $\{\hat{i}, j\}$ *and* h *are independent given* g*:* $p(\hat{i}, j, h|g) = p(\hat{i}, j|g)p(h|g)$.

We now rewrite this using expressions that follow from the definition for conditional probabilities. The left-hand side is written as $p(\hat{i}, j, h|g) = p(j, h|\hat{i}, g)p(\hat{i}|g)$*. The right-hand side can be rewritten using* $p(\hat{i}, j|g) = p(j|\hat{i}, g)p(\hat{i}|g)$.

This finally leads the conclusion that $p(j, |\hat{i}, g, h) = p(j|\hat{i}, g)$*. This last expression also implies that* $p(j|\hat{i}, g)p(g|\hat{i}, h) = p(j, g|\hat{i}, h)$*. Both expressions will be used later in this article.*

*Now consider the situation depicted in Figure 1(**b**). According to the Causal Markov Condition* $\{\hat{i}, j\}$ *and* h *are* not *independent given* g*, i.e.,* $p(\hat{i}, j, h|g) \neq p(\hat{i}, j|g)p(h|g)$*. We can still rewrite the left-hand side and the right-hand side in the same fashion as before.*

This finally leads the conclusion that $p(j, |\hat{i}, g, h) \neq p(j|\hat{i}, g)$*. This implies that* $p(j|\hat{i}, g)p(g|\hat{i}, h) \neq p(j, g|\hat{i}, h)$.

In the example above we used a simplified notation for the probabilities $p(y) := Pr\{Y = y\}$. As illustrated in the above example we use basic aspects of probability theory like the definition of joint probabilities and the Law of Total Probability [22]. This law links a marginal probability to a joint probability, e.g., $\sum_g p(j, g|i, h) = p(j|i, h)$.

Unless stated otherwise, in this article we will make use of the Einstein summation convention (with a twist). This convention simplifies equations by implying summation over indices that appear both as upper indices and as lower indices. In our definition the summation takes place the first indices and the subsequent identical lower indices $B^i_j A^j_i := \sum_i B^i_j A^j_i$. Our definition implies that $B^i_j A^j_i \neq A^j_i B^i_j$, the order matters.

2. Information Theory

Shannon introduced information theory in 1948 [23]. It models association between random variables as resulting from a communication process between a sender—the source—and a receiver or the destination. A message comprises indexed realizations of random variables representing stationary ergodic processes. An input message is first encoded: we describe the message using a finite alphabet. Each random variable has its own finite alphabet. The random variable X is mapped on symbols from the alphabet $\mathcal{X}$, the random variable Y is mapped on symbols from $\mathcal{Y}$, and the random variable Z is mapped on symbols from $\mathcal{Z}$. Where $\mathcal{X} = \{\chi_1, \chi_2, \cdots, \chi_{|\mathcal{X}|}\}$, $\mathcal{Y} = \{\psi_1, \psi_2, \cdots, \psi_{|\mathcal{Y}|}\}$, and $\mathcal{Z} = \{\zeta_1, \zeta_2, \cdots, \zeta_{|\mathcal{Z}|}\}$. The number of elements in the alphabet—the cardinality—is denoted as $|\mathcal{X}|$, $|\mathcal{Y}|$, and $|\mathcal{Z}|$ respectively.

Once encoded the message is transmitted symbol by symbol. The input symbol is transformed into an output symbol. The output alphabet can have a different cardinality than the input alphabet. The transformation from input to output symbol is modeled as a Markov chain. The probability that a specific output symbol is received only depends on the alphabet symbol that was sent. The communication process transforms the input probability mass function (pmf) into the output pmf. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

transmitted message is decoded and made available to the receiver. In this article, we assume that no decoding takes place.

2.1. Mutual Information

If there is an association between two messages, information is said to be shared between them. The amount of information shared,

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2\left[\frac{p(y|x)}{p(y)}\right], \tag{1}$$

is nonnegative and symmetric in X and Y. This so-called mutual information (MI) represents the reduction in uncertainty about the random variable X given that we have knowledge about the random variable Y (and vice versa). It is intuitively clear that, given the information content of the source data, in subsequent transmissions, the information can never increase. This is formalized in the Data Processing Inequality or DPI which states that processing of data can never increase the amount of information [20]. For the cascade $X\rightarrow Y\rightarrow Z$ the DPI implies that, in terms of MI,

$$I(X;Z)\leq\min[I(X;Y),I(Y;Z)].$$

The maximum rate with which information can be transmitted between the sender and receiver is the channel capacity $C_{XY}=\max_{p(x)}[I(X;Y)]$. This is achieved for a so-called channel achieving input distribution.

2.2. The Communication Channel

In information theory, the directed graph representing a Markov chain is represented as a *communication channel*, or channel in short. The channel has an input side—the left-hand side—and an output side—the right-hand side. On the left-hand side we place all the vertices of the Markov chain with outgoing edges and on the right-hand side we place all the vertices of the Markov chain with incoming edges. The input vertices are connected to the output vertices via undirected edges. In a channel, every input alphabet symbol has its own input vertex. Likewise, every output alphabet symbol has its own output vertex. The simplest type of channel is the noisy discrete memoryless communication channel. In a memoryless channel the output—y_t—only depends on the input—x_t—and not on the past inputs or outputs: $p(y_t|x_t,x_{t-1},y_{t-1})=p(y_t|x_t)$. A memoryless channel embodies the Markov property. In a noisy channel the output depends on the input and another random variable representing noise. The effect of transmitting data using a DMC is described via the Law of Total Probability because

$$Pr\{Y=\psi_j\}=\sum_i Pr\{X=\chi_i\}Pr\{Y=\psi_j|X=\chi_i\}, \tag{2}$$

with $Pr\{Y=\psi_j\}$ the j^{th} element of the probability mass function $p(y)$, and $Pr\{X=\chi_i\}$ the i^{th} element of the pmf $p(x)$. The transmission of data over a DMC transforms the pmf of the input into the pmf of the output via a linear transformation. The probability transition matrix $Pr\{Y=\psi_j|X=\chi_i\}$ fully characterizes the DMC [20]. Assuming a fixed (e.g. lexicographic) order of the alphabet elements, we can introduce an index notation for the pmfs, e.g, $p^j:=Pr\{Y=\psi_j\}$ and $p^i:=Pr\{X=\chi_i\}$. In this article we associate every index with a specific random variable. In Table 1 an overview is given.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Table 1. Overview of Indices Used.

Process	Variable	Alphabet element	Index (input)	Index (past)	Index (output)
X	x	χ	$\hat{i}$	f	i
Y	y	ψ	$\hat{j}$	g	j
Z	z	ζ	$\hat{k}$	h	k

2.3. Tensor Representation of the Communication Channel

One of the many virtues of information theory is that it enables the use of linear algebra. Because we do not want to get overwhelmed by increasingly complex probabilistic equations, we use index notation and the Einstein summation convention (with a minor twist). Equation (2) can now be written as $p^j = p^i A_i^j$. The covariant indices indicate the variables we condition on. The row stochastic probability transition matrix elements represent the elements of the probability transition tensor A [26]. Using the standard notation instead of the Einstein summation convention, we can rewrite MI as $I(X;Y) = \sum_{i,j} p^{ij} \log_2 \left[A_i^j / p^j \right]$.

Mutual information solely depends on the elements of the tensor and the input pmf. This is problematic in case MI or MI derived measures are used to infer the underlying structure if we assume that the structure is independent of the input. We can illustrate this by assuming that the probability transition tensor equals the Kronecker delta

$$\delta_i^j = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Example 2. *If $A_i^j = \delta_i^j$, the symbol received is identical to the symbol sent, the channel transmits data perfectly. In this case MI reduces to $I(X;Y) = \sum_i p^i \log_2 \left[1/p^i \right]$. Now set the probability of one of the alphabet elements to $1 - \varepsilon$. This implies that all other symbol probabilities are equal to or smaller than ε. Taking the limit $\varepsilon \to 0$ results in a mutual information $\to 0$. Although there might be a noiseless channel representing the association between the random variables X and Y, MI could be arbitrarily small.*

This leads us to the following proposition for inferring structures using MI-based measures:

Proposition 1. *In case MI or MI related measures are used to infer the **structure** for a system, we should use the probability transition tensors or measures based on elements of probability transition tensor.*

The earlier mentioned channel capacity is such a measure. It only depends on the elements of the probability transition tensor [27], e.g. $C_{XY} := \Gamma(\mathsf{A})$. Because the channel capacity is the maximal achievable mutual information for a specific channel, the earlier mentioned DPI also applies to the channel capacity. The proof is straightforward and therefore omitted.

Corollary 1 (DPI for Channel Capacity)**.** *For the chain $X \to Y \to Z$ the DPI immediately implies that $\Gamma(\mathsf{C}) \leq \min[\Gamma(\mathsf{A}), \Gamma(\mathsf{B})]$, with A representing the tensor of the transmission $X \to Y$, $\mathsf{B}: Y \to Z$, and $\mathsf{C}: X \to Z$.*

In this short and incomplete introduction to information theory, no assumptions—other than stationarity, ergodicity and Markov property—were made about the underlying mechanisms leading to the association between random variables. We can therefore apply information theory to all cases where observational data are available.

3. Transfer Entropy

Schreiber introduced transfer entropy in 2000 [8]. Like MI it is non-parametric, but unlike MI it is an essentially asymmetric measure and it enables the differentiation between a source and a

destination. It is an information theoretical implementation of Wieners principle of Causality [28]: a cause combined with the past of the effect predicts the effect better than that the effect predicts itself. In contrast to Granger causality [18], transfer entropy can capture nonlinear relationships. We use a slightly modified version which was shown to comply fully with Wieners principle of Causality by Wibral et al. They proved that this modified TE is maximal for the real interaction delay [29]. We assume that Y is a Markov process of order $\ell \geq 1$. This implies that the future y_t also depends on its past $\mathbf{y}^- = (y_{t-1}, \cdots, y_{t-\ell})$. The destination also depends on the source data X. With τ the finite interaction delay, we assume that for the input symbol $\mathbf{x}^- = (x_{t-\tau}, \cdots, x_{t-\tau-m})$, with $m \geq 0$. The alphabet for the past of Y is $\mathcal{Y}^\ell$. The alphabet for the input is $\mathcal{X}^m$.

$$TE_{X\to Y} = \sum_{\substack{\mathbf{x}^- \in \mathcal{X}^m, y \in \mathcal{Y} \\ \mathbf{y}^- \in \mathcal{Y}^\ell}} p(\mathbf{x}^-, y, \mathbf{y}^-) \log_2 \left[\frac{p(y|\mathbf{x}^-, \mathbf{y}^-)}{p(y|\mathbf{y}^-)} \right] \tag{3}$$

To differentiate a source from a destination, we have to assess two hypotheses: (1) X is the source and Y is the destination, and (2) Y is the source and X is the destination. Per case the interaction delay that maximizes the respective TE is determined. If the resulting transfer entropy equals 0, we assume that there is no relation. If the TE values are larger than 0, there are in practice two possibilities: (1) the optimal interaction delays are equal: we assume that the hypothesis resulting in the largest TE is valid. (2) The optimal interaction delays are different: both hypotheses are valid so we have detected a cycle. Without loss of generality, we assume in this article that there are no cycles.

Transfer entropy is a conditional mutual information [8]. Therefore, it can be associated with communication channels. We start with conditioning the MI from Equation (1) on the event $\mathbf{y}^- = \psi_g^-$ resulting in

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

Because $\mathbf{x}^-$ and $\mathbf{y}^-$ are the only parents of the output y, it follows from the Causal Markov Condition that the associated channel is memoryless. This mutual information quantifies the amount of information that is transmitted over the g^{th} sub-channel. The transfer entropy from Equation (3) can now be expressed as

$$TE_{X\to Y} = \sum_{\psi_g^- \in \mathcal{Y}^\ell} p(\psi_g^-) I(X;Y|\psi_g^-). \tag{4}$$

We now show that this expression results from transmission of information over a set of communication channels.

3.1. The Causal Channel

An inverse multiplexer comprises a demultiplexer and a multiplexer in series. A demultiplexer separates an input data stream into multiple output data streams. We call these different streams sub-channels. A multiplexer combines or multiplexes all input data streams into a single output data stream [30].

Definition 1 (Causal Channel). *A causal channel is an inverse multiplexer in which the demultiplexer selects the sub-channel over which the data are send based on the past of the output data. Each sub-channel consists of a DMC. The input symbol is fed to a specific input vertex of the chosen discrete memoryless channel. The DMC transforms the input in a probabilistic fashion into an output symbol. The multiplexer combines the outputted symbols into the output message. See Figure 2(**a**).*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

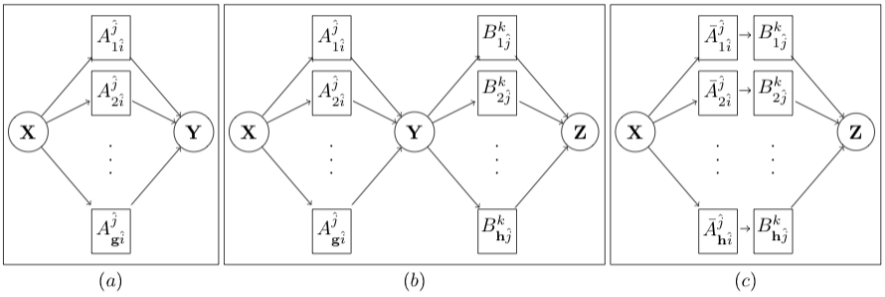

Figure 2. (**a**) Causal channel. (**b**) Two causal channels in series representing the communication model related to transfer entropy for the cascade $X \to Y \to Z$. (**c**) The equivalent causal channel for two causal channels in series.

This definition forms the basis for the theorem that is central to this article.

Theorem 1. *Transfer entropy is the average conditional mutual information of transmission over a causal channel.*

Proof. The relative frequency with which the g^{th} sub-channel is chosen equals $p(\psi_g^-)$. Each sub-channel is a DMC, so the mutual information of the g^{th} sub-channel equals $I(X;Y|\psi_g^-)$. The weighted average of the mutual information over all the sub-channels is equal to $\sum_{\psi_g^- \in \mathcal{Y}^\ell} p(\psi_g^-) I(X;Y|\psi_g^-)$, which is the definition of TE in Equation (4). □

Because a DMC is a causal channel with only one sub-channel, we call a DMC a *single-channel causal channel*.

3.2. Tensor Representation of a Causal Channel

Because every sub-channel of the causal channel represents a DMC, a causal channel can be represented by a probability transition tensor. We will call this tensor a *causal tensor*. For the relation $X \to Y$ we get the following equation for the g^{th} sub-channel

$$p_g^j = p_g^{\hat{i}} A_{g\hat{i}}^j. \tag{5}$$

The elements of the tensor A are given by $A_{g\hat{i}}^j = p(\psi_j|\chi_{\hat{i}}^-, \psi_g^-)$. We can now rewrite TE as

$$TE_{X\to Y} = \sum_{g,\hat{i},j} p^{\hat{i}jg} \log_2 \left[\frac{A_{g\hat{i}}^j}{p_g^j} \right]. \tag{6}$$

In a similar fashion as MI, we can show that TE can be made arbitrarily close to 0 while the causal tensor itself represents a noiseless transmission. It is therefore not an optimal measure to infer structures. Again we would prefer to use the tensors themselves or measures based on these tensors like the channel capacity. The determination of the channel capacity for a causal channel is not in the scope of this article. We assume however that it is possible to determine the channel capacity.

As stated in the introduction, the approach in the article was inspired by Turing machines. The causal tensor is a realization of the transition function of a Turing machine that encodes causality in as far as the causality is encoded in the pmfs. To warrant the use of the adjective "causal" however, we have to show that within the framework of causal tensors; we can differentiate between direct and indirect associations. That this seems possible can be intuited when considering the chain $X \to Y \to Z$ (see Figure 3a). The relation $X \to Z$ is a resultant of the other relations, i.e., an indirect association.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Within the framework of causal tensors, we would expect that we can express this indirect association in terms of the tensors of the other relations.

4. Calculation Rules for Causal Tensors

In this section we derive the calculation rules for causal tensors. Operations performed on these tensors should result in either scalars, stochastic vectors or stochastic tensors. To prove that we end up with stochastic tensors or vectors, we will use the earlier introduced index notation, the law of total probability and the Causal Markov Condition. We derive the calculation rules by investigating the four elementary structures depicted in Figures 3(**a**), 3(**b**), 3(**c**) and 3(**d**).

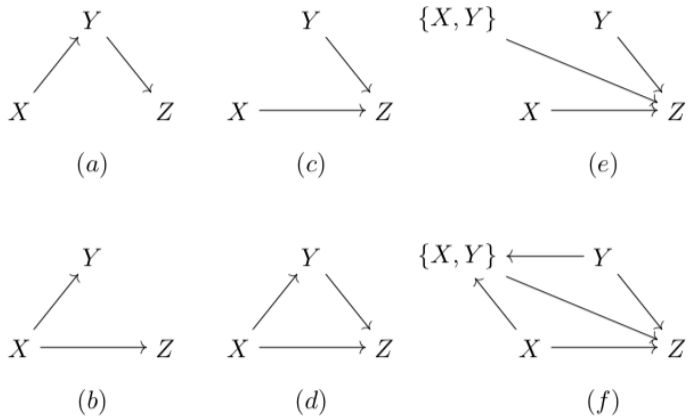

Figure 3. The basic structures directed graph structures: (**a**) the chain, (**b**) the fork, (**c**) the v-structure, and (**d**) the directed triangle. The graphs (**e**) and (**f**) reflect the calculation rules for the causal tensors for the v-structure and directed triangle respectively.

In Section 5 we will propose another notation which is simpler, but it relies on the results in this section.

4.1. The Chain Structure

First assume the chain $X \to Y \to Z$ is the ground truth. Additional to Equation (5), $p_g^j = p_g^{\hat{i}} A_{g\hat{i}}^j$, there are two other causal channels represented by tensors: $\mathsf{B} : Y \to Z$, and $\mathsf{C} : X \to Z$. Because it is a straightforward exercise in which we again make use of the law of total probability, we leave it to the reader to confirm that

$$p_h^k = p_h^{\hat{j}} B_{h\hat{j}}^k, \tag{7a}$$

$$p_h^k = p_h^{\hat{i}'} C_{h\hat{i}'}^k. \tag{7b}$$

In principle the index $\hat{i}'$ in Equation (7b) is the index representing a different input vector than the index $\hat{i}$ in Equation (5); although they both refer to the random variable X. This is because $\hat{i}'$ is related to the source $\mathbf{x}'^- \in \mathcal{X}^{m'}$ of Z and $\hat{i}$ is the index related to the source $\mathbf{x}^- \in \mathcal{X}^m$ of Y. The Markov property however immediately implies that we can use one and the same index in both cases if we select the source vector with the largest cardinality. Without loss of generality, we use $\hat{i}$.

We can express the tensor of the indirect relation in terms of the tensors of the direct relations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Theorem 2 (Product Rule for a Chain). *Let* A *and* B *be the causal tensors of two causal channels in series. Let the tensor* C *represent the resulting indirect causal channel that must be measured in a bivariate approach. The tensor elements of* C *are given by*

$$C^k_{h\hat{\imath}} = p^g_{h\hat{\imath}} A^{\hat{\jmath}}_{g\hat{\imath}} B^k_{h\hat{\jmath}}. \tag{8}$$

If the ground truth is a directed triangle, the product rule for a chain is invalid, i.e., $C^k_{h\hat{\imath}} \neq p^g_{h\hat{\imath}} A^{\hat{\jmath}}_{g\hat{\imath}} B^k_{h\hat{\jmath}}$.

The proof is given in Appendix A.1. The term $p^g_{h\hat{\imath}} A^{\hat{\jmath}}_{g\hat{\imath}}$ is very interesting. In Appendix A.2 we proof that it is a stochastic tensor.

Lemma 1. *For a chain the product* $p^g_{h\hat{\imath}} A^{\hat{\jmath}}_{g\hat{\imath}}$ *is a stochastic tensor* $\bar{A}^{\hat{\jmath}}_{h\hat{\imath}}$.

The causal tensor $p^g_{h\hat{\imath}} A^{\hat{\jmath}}_{g\hat{\imath}}$ can be interpreted as the weighted average of the causal tensor A, given the h^{th} sub-channel of the final causal channel and the input $\hat{\imath}$. We can now rewrite Equation (8) as

$$C^k_{h\hat{\imath}} = \bar{A}^{\hat{\jmath}}_{h\hat{\imath}} B^k_{h\hat{\jmath}}. \tag{9}$$

If both A and B represent discrete memoryless channels we get the simpler, well known, product rule for a chain of DMC's.

Corollary 2 (Product Rule for a Chain of DMC's). *Let* A *and* B *be the causal tensors of two DMC's in series and let the tensor* C *represent the resulting, indirect, causal channel that must be measured in a bivariate approach. The tensor elements of* C *are given by*

$$C^k_{\hat{\imath}} = A^{\hat{\jmath}}_{\hat{\imath}} B^k_{\hat{\jmath}}. \tag{10}$$

The proof follows directly from the definition of a DMC in terms of a causal channel: a discrete memoryless channel is a causal channel comprising only one sub-channel, i.e., it is a single-channel causal channel. Combined with Lemma 1, this corollary leads to a very specific interpretation of Equation (8). According to Corollary 2, Equation 9 can be interpreted as representing two DMC's in series for the h^{th} sub-channel. This means we have an alternative structure for two causal channels in series as depicted in Figure 2(**c**).

Because the Data Processing Inequality applies to a cascade of discrete memoryless channels, the alternative structure suggests that there is a DPI for transfer entropy. In Section 6 we show that this is indeed the case. If one so wishes we could check if the measured TE for the potential spurious association equals the expected TE. For the chain $X \to Y \to Z$ the expected $TE_{X\to Z}$ is given by

$$TE_{X\to Z} = \sum_{\hat{\imath},h,k} p^{\hat{\imath}hk} \log_2 \left[\frac{\sum_{\hat{\jmath}} \bar{A}^{\hat{\jmath}}_{h\hat{\imath}} B^k_{h\hat{\jmath}}}{p^k_h} \right].$$

4.2. The Fork Structure

In this section, we show that a fork can be interpreted as a chain. The product rule for a chain is therefore also applicable to a fork. Assume that the fork is the ground truth (Figure 3(**d**)). Again we want to express the indirect association represented by B in terms of the other causal tensors. First, we notice that the input distribution can be reconstructed from the output distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Definition 2 (Reconstruction Operator). *The ‡-operator, or reconstruction operator, reconstructs the source distribution, conditioned of the past of the destination, from the destination distribution, conditioned of the past of the destination:*

$$p^{\hat{i}}_{g} = p^{j}_{g} A^{\ddagger\hat{i}}_{gj}, \tag{11}$$

with $A^{\ddagger\hat{i}}_{gj} = p^{\hat{i}}_{gj}$. *The ‡-operation changes the sign of the interaction delay of the original relation.*

This implies that the directed graph $X \to Y$, is equivalent to the graph $X \leftarrow^{\ddagger} Y$. Because it is straightforward using Equation (6) we leave it to the reader to prove the following corollary.

Corollary 3. *From an information theory point of view, a relation and its reconstructed relation are equivalent, i.e.,* $TE_{X \to Y} = TE_{X \leftarrow \ddagger Y}$

From this corollary immediately follows that a fork is equivalent to a chain.

Theorem 3 (Fork-Chain Equivalence). *The fork* $X \to Y + X \to Z$ *is equivalent to the chain* $Y^{\ddagger} \to X \to Z$ *and to the chain* $Y \leftarrow X \leftarrow^{\ddagger} Z$.

The indirect association represented by B in terms of the other two tensors of the chain follows directly from the product rule for a chain (Theorem 2).

$$B^{k}_{h\hat{j}} = \bar{A}^{\ddagger\hat{i}}_{h\hat{j}} C^{k}_{h\hat{i}}, \text{ with } \bar{A}^{\ddagger\hat{i}}_{h\hat{j}} := p^{g}_{h\hat{j}} A^{\ddagger\hat{i}}_{g\hat{j}}, \tag{12a}$$

$$B^{j}_{g\hat{k}} = \bar{C}^{\ddagger\hat{i}}_{g\hat{k}} A^{j}_{g\hat{i}}, \text{ with } \bar{C}^{\ddagger\hat{i}}_{g\hat{k}} := p^{h}_{g\hat{k}} C^{\ddagger\hat{i}}_{h\hat{k}}. \tag{12b}$$

Equation (12a) applies in case the equivalent chain is $Y \to^{\ddagger} X \to Z$. If the equivalent chain is $Y \leftarrow X \leftarrow^{\ddagger} Z$, Equation (12b) is applicable. Due to the way we determine the interaction delay, the ‡-operation induces a sign change for the interaction delay. E.g., if τ_{xy} represents the interaction delay for the relation $X \to Y$, then $-\tau_{xy}$ represents the interaction delay for the relation $Y^{\ddagger} \to X$.

Matrices, and therefore tensors, do not commute: $\bar{A}^{\ddagger\hat{i}}_{h\hat{j}} \bar{A}^{\hat{j}}_{h\hat{i}} \neq \bar{A}^{\hat{j}}_{h\hat{i}} \bar{A}^{\ddagger\hat{i}}_{h\hat{j}}$. This leads to the conclusion that a chain and a fork are in principle distinguishable. The reader can verify this by combining Equation (9) and Equation (12a). Combining this with Theorem 2 leads to the conclusion that we can differentiate between direct and indirect relations. The conditions under which it is not possible will be derived later.

Theorem 4. *Using causal tensors we can differentiate between a chain, a fork and a directed triangle. If and only if the chain is the ground truth* $B^{k}_{h\hat{j}} \neq \bar{A}^{\ddagger\hat{i}}_{h\hat{j}} C^{k}_{h\hat{i}}$. *If and only the fork is the ground truth* $C^{k}_{h\hat{i}} \neq \bar{A}^{\hat{j}}_{h\hat{i}} B^{k}_{h\hat{j}}$. *If the structure is neither a chain, nor a fork, it is a directed triangle.*

In case of a single-channel causal channels, we are allowed to ignore the index indicating the sub-channels. The equations for the chain, the fork, and consequently the equations in Theorem 4 are all bivariate.

Corollary 4. *If we use single-channel causal tensors, bivariate time-delayed mutual information measurements can differentiate between a fork, a chain, and a directed triangle.*

This result contradicts the current point of view [12,19]. We illustrate this with an example. Because we use single-channel causal channels, we ignore the index h in the equations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Example 3. *Let the chain $X \to Y \to Z$ be the ground truth. With* $\mathsf{A} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$ *and* $\mathsf{B} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ 0 & 1 \end{pmatrix}$*, the indirect association is represented by the causal tensor* $C^k_{\hat{i}} = A^{\hat{j}}_{\hat{i}} B^k_{\hat{j}} \Rightarrow \mathsf{C} = \begin{pmatrix} \frac{1}{6} & \frac{5}{6} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$*. Assume that* $p(x) = (\frac{2}{5}, \frac{3}{5})$*. The pmf for* $p(y)$ *equals* $p(y) = p(x)\mathsf{A}$*. From this follows that* $p(y) = (\frac{4}{5}, \frac{1}{5})$*. Using the relation* $p(x) = p(y)\mathsf{A}^\ddagger$*, the reader can verify that* $\mathsf{A}^\ddagger = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ 1 & 0 \end{pmatrix}$*. Because* $\bar{A}^{\ddagger\hat{i}}_{h\hat{j}} C^k_{h\hat{i}} = \begin{pmatrix} \frac{7}{24} & \frac{17}{24} \\ \frac{1}{6} & \frac{5}{6} \end{pmatrix}$*. From this follows that* $B^k_{h\hat{j}} \neq \bar{A}^{\ddagger\hat{i}}_{h\hat{j}} C^k_{h\hat{i}}$*. The structure is that of a chain, and not that of a fork.*

4.3. The V-Structure and the Directed Triangle

In a bivariate measurement, we will always be able to determine the ground truth correctly in the case of the v-structure depicted in Figure 3(**c**). However, investigating structures with a collider, the v-structure and the more general directed triangle, will result in the important concept of *interaction*. So, let us assume that the ground truth is the directed triangle. We now have to introduce the multivariate relation $\mathsf{D} : \{X, Y\} \to Z$. This relation leads to the additional linear transformation

$$p^k_h = p^{\hat{i}\hat{j}}_h D^k_{h\hat{i}\hat{j}}.$$

We call the tensor D the *interaction tensor*. The tensors B and C can be expressed in terms of the tensor D.

Lemma 2 (Causal Tensor Contraction)**.** *In the case of a directed triangle, we can express the causal tensors in terms of the interaction tensor:*

$$B^k_{h\hat{j}} = \bar{A}^{\ddagger\hat{i}}_{h\hat{j}} D^k_{h\hat{i}\hat{j}}, \tag{13a}$$

$$C^k_{h\hat{i}} = \bar{A}^{\hat{j}}_{h\hat{i}} D^k_{h\hat{i}\hat{j}}. \tag{13b}$$

For the proof, we use the fact that the elements of a causal tensor are conditional probabilities. Due to the fork-chain equivalence, Appendix A.3 only contains the proof for the chain.

From Equation (13) it follows that B and C are the result of a cascade involving $\mathsf{A}^\ddagger$ and D for B, and A and D for C. The graphs represented by Figures 3(**e**) and 3(**f**) support the tensor relations, $X \to \{X, Y\} \to Z$ is equivalent to the cascade of the inverse multiplexers represented by A and D resulting in C. Figures 3(**c**) and 3(**d**) however do not support the calculation rules for causal tensors.

Proposition 2. *If a complex system contains v-structures, the causal graph must be represented by a directed hypergraph [31]. In a hypergraph, an edge connects any number of vertices. The interaction tensor corresponds to a so-called hyperedge.*

The interaction tensor describes the interaction of inputs at the v-structure. An indirect relation does not interact.

Theorem 5. *The interaction tensor only depends on the direct causes, not on indirect causes. So, if and only if the chain is the ground truth* $D^k_{h\hat{i}\hat{j}} = B^k_{h\hat{j}}$*. If and only if the fork is the ground truth* $D^k_{h\hat{i}\hat{j}} = C^k_{h\hat{i}}$*.*

Sketch of Proof. Let the ground truth be the chain. In that case, $X \perp\!\!\!\perp Z | Y$ and X is a non-effect of Z. The index $\hat{i}$ is associated with X, the index $\hat{j}$ is associated with Y and the indices h and k are associated with Z. The Causal Markov Condition leads to $p^{\hat{i}k}_{h\hat{j}} = p^k_{h\hat{j}} p^{\hat{i}}_{h\hat{j}} \Leftrightarrow p^k_{h\hat{i}\hat{j}} = p^k_{h\hat{j}}$. □

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

In this index ridden section, we have shown that we can express indirect relations in terms of the direct relations; the resulting tensors are stochastic tensors. We can now introduce a notation that simplifies our expressions.

5. A New Notation

The input for a causal tensor is a probability vector. This vector represents the *probabilistic state* of the input. Let us assume that the input and output alphabets each consists of five alphabet elements. Because we place the operand before the operator, the probability vectors are row vectors. An input event x could be represented by the probability vector $(0,0,1,0,0)$, i.e., the input equals the 3rd alphabet element. After the transmission over the causal channel, the probabilistic *state* of the output element could, for example, be $(0,0.3,0,0.2,0.5)$, a probabilistic mix of the output alphabet elements.

Apart from the probabilistic state, the states of the input and output events should also accommodate the *temporal* relationship between the input and output, indicated by the interaction delay. We can unify both state characteristics by introducing complex state vectors and by adopting the *braket* or Dirac notation used in quantum mechanics [32].

5.1. Braket Notation and Tensor Operations

The complex state vector, or the *ket* is represented by the notation $|x\rangle$. From the viewpoint of index notation, the order of the operand and operator is irrelevant, so we chose to represent the probability component as a column vector, i.e., the transpose of the probability vector. In the case of our example this would be $(0,0,1,0,0)^T$. The temporal dependencies are taken care of via a complex number, e.g., $e^{i\omega t_x}$. The i in $e^{i\omega t_x}$ is not an index, but the imaginary number i. The ω represents a normalizing constant and t_x a time constant time related to the random variable, in this case, X. Each random variable has its own constant time associated with it. The ket in our example is represented by $|x\rangle = (0,0,1,0,0)^T e^{i\omega t_x}$.

We assume that the interaction delay is induced by the causal channel. We therefore associate it with the causal tensor. We represented it by $e^{i\omega\tau_{xy}}$, where τ_{xy} is the interaction delay between X and Y. The *complex* causal tensor is defined as $\mathcal{A} := \mathsf{A}^T e^{i\omega\tau_{xy}}$. We can now rewrite Equation (5) as $|y\rangle = \mathcal{A}\,|x\rangle$. In the case of our example, the ket for the output is given by $|y\rangle = (0,0.3,0,0.2,0.5)^T e^{i\omega(t_x+\tau_{xy})}$. To be able to simplify the product rule for a chain given by Equation (8), we define the cascading operator using the chain $X \to Y \to Z$.

Definition 3. *The causal tensor cascading operator $\odot$ applied to a cascade of two causal tensors, $\mathcal{A}$ and $\mathcal{B}$, is defined as*

$$\mathcal{B} \odot \mathcal{A}\,|x\rangle = |z\rangle\,.$$

Equation (9) can now be written as

$$\mathcal{C} = \mathcal{B} \odot \mathcal{A}. \tag{14}$$

Our definition implies that the interaction delay of a cascade of causal channels is the sum of all the interaction delays between the subsequent pairs making up the cascade. This is not an uncommon assumption [33,34]. The additivity also applies to "daggered" relations.

Definition 4.

$$\mathcal{A}^\ddagger := (\mathsf{A}^\ddagger)^T e^{-i\omega\tau_{xy}}.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Apart from the fact that index ridden equations can be simplified, and independent of our historical attachment as the framework resulted from the intuition that a notation like this should be possible, it is a subject of future research if this novel notation provides additional new insights.

5.2. Transfer Entropy and Mutual Information

Both TE and MI can be expressed in terms of the complex tensors. First, we introduce a simplified notation for TE We write these measures as a function of pmfs, indicated by $(\cdot)$ and the respective tensor.

Corollary 5.

$$TE_{X\to Y} = \|TE(\mathcal{A},\cdot)\|,$$

with $\|\cdot\|$ *the Euclidean norm.*

The proof is straightforward when using Equation (6) and we therefore omit it. From now on we always assume that for measures like TE and MI we have to take the Euclidian norm. This allows us to write $TE_{X\to Y} = TE(\mathcal{A},\cdot)$. We use the subscripts, e.g., h, to indicate the slice representing the sub-channel in a multi-channel causal channel, e.g., $I(\mathcal{A}_h,\cdot)$ represents the MI for the h^{th} sub-channel.

6. Inferring Structures With Causal Tensors

In this section, we discuss some non-trivial implications when using causal tensors to infer the causal structure from time-series data. First, we will show that a Data Processing Inequality for transfer entropy exists. Because we did not make any assumption about the cardinality of the alphabets used, this DPI is also valid for time-discrete, continuous data. We then prove that we can differentiate between a fork, a chain and a directed triangle as long as the data are noisy, but not "perfectly noisy"—we will define later this in this article. Finally we will establish a theorem that enables us to always use bivariate analysis within a system comprising three variables.

6.1. Data Processing Inequality for TE

The DPI for TE gives a sufficient condition to assess if a relation is a proper direct relation. It gives a necessary condition to detect potential indirect relations.

Theorem 6 (DPI for a Chain). *For the chain* $X\to Y\to Z$ *the following inequality holds*

$$TE_{X\to Z} \leq \min\left[TE_{X\to Y}, TE_{Y\to Z}\right]. \tag{15}$$

Because the fork has equivalent chains, the DPI also applies to a fork.

For readability, we have moved the proof to Appendix A.4. Under the condition that the embedding of the cause and the effect vectors is sufficiently large, the DPI can identify potential indirect relations. Because we made no assumptions about the cardinality of the (finite) alphabets, the DPI is also valid for finite, very large alphabets.

6.2. Differentiating Between Direct and Indirect Associations With Causal Tensors

We have shown earlier that a fork, a chain, and a directed triangle are distinguishable. We now investigate in more detail under what conditions this is not possible.

Definition 5 (Perfect Noisy Relation). *If and only if all causal tensor elements are equal, the relation is a perfect noisy relation. The related causal tensor is called the perfect noisy causal tensor.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

The behavior of a perfect noisy causal tensor is straightforward and therefore left to the reader to confirm: (1) any input pmf is transformed into a uniform probability distribution, (2) the channel capacity = 0. The opposite of the perfect noisy causal tensor is the noiseless causal tensor.

Definition 6 (Noiseless Causal Tensor). *The elements of a noiseless causal tensor satisfy* $\forall_{h\hat{i}\hat{j}} A^{\hat{j}}_{h\hat{i}} \in \{0,1\} \cup \forall_h : \sum_{\hat{i}} A^{\hat{j}}_{h\hat{i}} = 1$ *and* $\sum_{\hat{j}} A^{\hat{j}}_{h\hat{i}} = 1$.

The reader can verify by using Equation (6) that for any input pmf, $TE = \log_2 \left[\sum_{\hat{j}} 1\right]$. Because the channel capacity of a noiseless channel only depends on the number of alphabet elements, $C_{XY} = \min \left[\log_2(|\mathcal{X}^m|), \log_2(|\mathcal{Y}^\ell|)\right]$ [20], our definition is indeed a noiseless causal channel. An immediate consequence of the definition of a noiseless tensor is that the cardinality of the input pmf equals the cardinality of the output pmf.

That we can differentiate between direct and indirect relations is related to noise. We proved the following theorem in Appendix A.5.

Theorem 7. *We cannot differentiate between direct and indirect relations if: (1) all relations are perfectly noiseless, or (2) the relations are (almost) perfectly noisy.*

6.3. Bivariate Analysis With Causal tensors

We now show that within our simple system of three variables, bivariate analysis suffices. We first need to determine the causal tensors representing multi-channel causal channels, after which we determine the causal tensors representing single-channel causal channels.

Theorem 8. *If in a system comprising three variables, the ground truth is a chain, then the product rule for a chain is applicable to both the multi-channel causal channels and the single-channel causal channels:*

$$C^k_{h\hat{i}} = \bar{A}^{\hat{j}}_{h\hat{i}} B^k_{h\hat{j}} \Leftrightarrow C^k_{\hat{i}} = A^{\hat{j}}_{\hat{i}} B^k_{\hat{j}}. \tag{16}$$

The proof can be found in Section A.6.

From the proof it immediately follows that this theorem is only valid when we use the same embedding and the same interaction delays for both the single-channel causal channels as for the multi-channel causal channels.

6.4. Causal Inference Steps

To complete the causal tensor framework as discussed so far, a summary of the (implicitly) proposed steps are given. We assume that: (1) the data are time equidistant, (2) ℓ and m are determined correctly, and (3) the data are ergodic and stationary.

1. Encode the data into a finite alphabet.
2. Determine the (bivariate) multi-channel causal tensors for a range of interaction delays.
3. Determine the optimal interaction delay.
4. Determine per relation the direction of causation.
5. Identify the potential indirect relations using the DPI and the additivity of interaction delays.
6. Determine the single-channel causal tensors for the potential indirect relations.
7. Use the product rule to determine if the indirect relations are indeed indirect.
8. If the network is used for simulation, determine the interaction tensors for all v-structures, i.e., determine the hypergraph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

7. Experiments

We finalize this article with two experiments to illustrate that nonlinear behavior is indeed captured with causal tensors.

7.1. Ulam Map

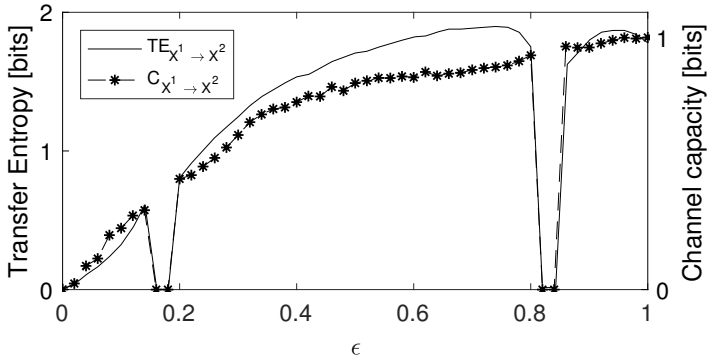

Figure 4. Transfer entropy and the channel capacity of the causal tensor for two unidirectionally coupled Ulam maps X^1 and X^2 as a function of the coupling strength ϵ. Only the relation $X^1 \to X^2$ is shown. Dots: approximated channel capacity for the causal channel. Line: transfer entropy as determined by Schreiber.

For the first experiment, we use the one-dimensional lattice of unidirectional coupled maps $x^m_{n+1} = f\left(\epsilon x^{m-1}_n + (1-\epsilon)x^m_n\right)$. Information can only be transferred from X^{m-1} to X^m. The Ulam map with $f(x) = 2 - x^2$ is interesting because there are two regions ($\epsilon \approx 0.18$, $\epsilon \approx 0.82$) where no information is shared between maps [8]. We chose an alphabet comprising four symbols. The quantization comprised simple binning. Furthermore we chose $\ell = m = 1$ (see Equation (3)). Instead of maximizing TE we maximized the channel capacity to determine the optimal delay. An approximation that satisfies the boundaries that follow from Equation (4) was used,

$$\tilde{\Gamma}(\mathsf{A}) = \sum_g p(\psi^-_g)\Gamma(\mathsf{A}_g). \tag{17}$$

To determine the channel capacities the Blahut-Arimoto algorithm was used [35]. The delays were varied between one and 20. The Channel capacity was maximal for a delay of one sample. As seen from Figure 4, causal tensors lead to a similar result as transfer entropy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

7.2. Coupled Ornstein-Uhlenbeck Processes

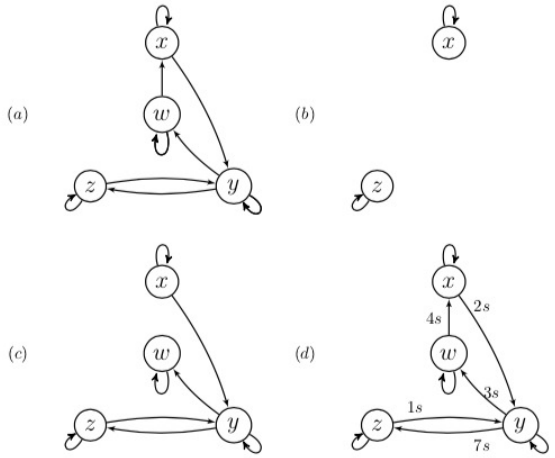

Figure 5. (**a**) The causal structure for the Ornstein-Uhlenbeck system of Equation (18). The other graphs show the inferred causal structures at different time series lengths. The confidence interval was 90% and the maximum delay was set to 20s: (**b**) $T = 10ks$, (**c**) $T = 100ks$. In (**d**), $T = 500ks$, the interaction delays that maximized the channel capacity are also shown.

In the second experiment, we demonstrate our approach using a system of four coupled Ornstein-Uhlenbeck processes [11]:

$$\begin{aligned}
\dot{x}(t) &= -0.5x(t) + 0.6w(t-4)\eta_x(t),\\
\dot{y}(t) &= -0.9y(t) - 1.0x(t-2) + 0.6z(t-5) + \eta_y(t),\\
\dot{z}(t) &= -0.7z(t) - 0.5y(t-6) + \eta_z(t),\\
\dot{w}(t) &= -0.8w(t) - 0.4y(t-3)^2 + 0.05y(t-3) + \eta_w(t),
\end{aligned} \tag{18}$$

with independent unit variance white noise processes η. The integration time step was $dt = 0.01s$ and the sampling interval $\Delta s = 100s$. We used a binary encoding scheme. First, the data was normalized after which it was partitioned at 0.5. Because the Shannon entropy of the encoded data was close to 1, we expect highly noisy communication channels. The disadvantage of binary encoding is that more data is needed to capture the transmitted information. However, cascading very noisy channels reduces the probability of detecting an indirect relation. This is illustrated by Figure 5, no pruning was needed. This experiment shows that causal tensors can indeed detect the underlying structure.

8. Discussion

In this article we focused mainly on the basic foundations of our formalism, we assumed for example that there were no hidden causes. Whether our formalism can be used to detect the existence of hidden causes will be researched in the future. If and how our framework applies to systems comprising over three random variables will also be a subject of further research.

The conclusion that bivariate analysis suffices to differentiate direct and indirect associations, follows directly from the DPI. It should not be of any surprise that in an information theoretical framework, noise plays a fundamental role. A cascade of noisy channels is at least as noisy as the nosiest channel within the cascade.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

Further experiments are needed to confirm that a combination of multi-channel causal channels and single-channel causal channels could be used for pruning inferred networks.

The experiment in Subsection 7.1 illustrated that causal tensors give similar results as transfer entropy. The experiment in Subsection 7.2 illustrated that our formalism can indeed infer the underlying structure. All relations that should have been found based on the differential equations were found. All interaction delays except the interaction delay for $Z \to Y$ were close or equal to the interaction delays in the differential equations. In both cases, we used a rather simple encoding scheme. This raises the question if and under what conditions an optimal encoding scheme exists that minimizes the cardinality of the alphabets, but keeps a sufficiently large amount of "causal information".

We will focus future research on the extension and application of the framework. Because in the case of a directed triangle the tensor $\bar{\mathrm{A}}$ is related to the redundancy, the *potential* relation to "Partial Information Decomposition" [36] will be further explored in the future.

Funding: This research received no external funding.

Acknowledgments: I would like to thank Errol Zalmijn for introducing me to the wonderful topic of transfer entropy and Marcel Brunt for helping me to implement our approach in Matlab. Also, thanks to Hans Onvlee, S. Kolumban, Rui M. Castro and T. Heskes for their comments on earlier versions of the manuscript. ASML PI System Diagnostics supported part of the work.

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

Abbreviations

The following abbreviations are used in this manuscript:

DMC	Discrete memoryless communication channel
MI	Mutual information
pmf	probability mass function
TE	Transfer entropy

Appendix A. Proofs

Appendix A.1. Theorem 2

For the proof of Theorem 2 we need to introduce two lemma's.

Lemma A1. $\forall\, g:\ B^k_{ghj} = B^k_{hj}$.

Sketch of Proof for Lemma A1. Another direct consequence of the Markov property is related to indices associated with the same random variable. As long as the index related to the past of the output—g—and the index related to the output—j—appear in the same tensor we are allowed to replace the output index by the input index. In our example, this means that we may replace j by $\hat{j}$ as long as we ensure that $\psi^-_{\hat{j}} = \{\psi_j, \psi^-_g\}$. This is always possible because of the Markov property: we either enlarge the cardinality of $\psi^-_{\hat{j}}$ or ψ^-_g. □

The next lemma follows directly from Example 1.

Lemma A2. *For the chain* $X \to Y \to Z$ *we have* $\mathcal{A}^{\hat{j}}_{\hat{i}gh} = \mathcal{A}^{\hat{j}}_{\hat{i}g}$.

Sketch of Proof for Lemma A2. Figure 1(**a**) depicts the situation of the chain $X \to Y \to Z$. According to the Causal Markov Condition $\{\hat{i}, j\}$ and h are independent given g, i.e., $p(\hat{i}, j, h|g) = p(\hat{i}, j|g)p(h|g)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

We now rewrite this using expressions that follow from the definition for conditional probabilities. The left-hand side is written as $p(\hat{i},j,h|g) = p(j,h|\hat{i},g)p(\hat{i}|g)$. The right-hand side can be rewritten using $p(\hat{i},j|g) = p(j|\hat{i},g)p(\hat{i}|g)$.

This finally leads the conclusion that $p(j,|\hat{i},g,h) = p(j|\hat{i},g)$, i.e., $\mathcal{A}^{\hat{j}}_{\hat{i}gh} = \mathcal{A}^{\hat{j}}_{\hat{i}g}$. □

Sketch of Proof for Theorem 2. Because of the Law of Total Probability we are allowed to condition Equation (5) on h and both Equation (7a) and Equation (7b) on g. This leads to

$$p^{\hat{j}}_{gh} = p^{\hat{i}}_{gh} A^{\hat{j}}_{ghi\hat{}}, \tag{A1a}$$

$$p^{k}_{gh} = p^{\hat{j}}_{gh} B^{k}_{ghj\hat{}}, \tag{A1b}$$

$$p^{k}_{gh} = p^{\hat{i}}_{gh} C^{k}_{ghi\hat{}}. \tag{A1c}$$

Substituting the expression for $p^{\hat{j}}_{gh}$ of Equation (A1a) in Equation (A1b) and combining the result with Equation (A1c) gives us $C^{k}_{ghi\hat{}} = A^{\hat{j}}_{ghi} B^{k}_{ghj\hat{}}$. Using Lemma(A1) and Lemma(A2) this can be rewritten as

$$C^{k}_{ghi\hat{}} = A^{\hat{j}}_{gi\hat{}} B^{k}_{hj\hat{}}. \tag{A2}$$

Finally, we multiply both sides with $p^{g}_{hi\hat{}}$. As the reader can confirm, the term $p^{g}_{hi\hat{}} C^{k}_{ghi\hat{}}$ equals $C^{k}_{hi\hat{}}$. This finally leads to Equation (8).

For the second part of the theorem, we refer to Figure 1(**b**). It depicts the situation of the directed triangle $X \to Y \to Z + X \to Z$. According to the Causal Markov Condition $\{\hat{i},j\}$ and h are *not* independent given g: $p(\hat{i},j,h|g) \neq p(\hat{i},j|g)p(h|g)$.

We now rewrite this using expressions that follow from the definition for conditional probabilities. The left-hand side is written as $p(\hat{i},j,h|g) = p(j,h|\hat{i},g)p(\hat{i}|g)$. The right-hand side can be rewritten using $p(\hat{i},j|g) = p(j|\hat{i},g)p(\hat{i}|g)$. □

Appendix A.2. Lemma 1

Sketch of Proof for Lemma 1. By definition $p^{g}_{hi\hat{}} A^{\hat{j}}_{gi\hat{}} = \sum_g p(g|h,\hat{i})p(\hat{j}|g,\hat{i})$. From Example 1(**b**) it follows that $\sum_g p(g|h,\hat{i})p(\hat{j}|g,\hat{i}) = \sum_g p(\hat{j},g|\hat{i},h)$.

Applying the law of total probability to the righthand side gives us $\sum_g p(\hat{j},g|\hat{i},h) = p(\hat{j}|\hat{i},h)$. In other words: $p^{g}_{hi\hat{}} A^{\hat{j}}_{gi\hat{}} = p(\hat{j}|\hat{i},h)$. □

Appendix A.3. Lemma 2

Sketch of Proof for Lemma 2. First we note that $p^{\hat{i}\hat{j}}_{h} = \delta^{\hat{j}'}_{\hat{j}} p^{\hat{j}}_{h} p^{\hat{i}}_{hj\hat{}'}$. Equation (4.3) can therefore be is rewritten as $p^{k}_{h} = \delta^{\hat{j}'}_{\hat{j}} p^{\hat{j}}_{h} p^{\hat{i}}_{hj\hat{}'} D^{k}_{hi\hat{}\hat{j}}$. Because we are allowed to change the order of $\delta^{\hat{j}'}_{\hat{j}}$ and $p^{\hat{j}}_{h}$ we get $p^{k}_{h} = p^{\hat{j}}_{h}\left(\delta^{\hat{j}'}_{\hat{j}} p^{\hat{i}}_{hj\hat{}'} D^{k}_{hi\hat{}\hat{j}}\right)$. Combining this with Equation (7a) results in an expression for $B^{k}_{hj\hat{}}$: $B^{k}_{hj\hat{}} = \delta^{\hat{j}'}_{\hat{j}} p^{\hat{i}}_{hj\hat{}'} D^{k}_{hi\hat{}\hat{j}}$. Because $\delta^{\hat{j}'}_{\hat{j}} p^{\hat{i}}_{hj\hat{}'} = p^{\hat{i}}_{hj\hat{}}$ we get Equation (13a). □

Appendix A.4. Theorem 6

For the proof of the data processing inequality, Theorem 6, the simplified notation for transfer entropy and mutual information from Section 5 is used: $TE_{X\to Y} := TE(\mathcal{A},\cdot)$, $TE_{Y\to Z} :=$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

$TE(\mathcal{B},\cdot)$, $TE_{X\to Z} := TE(\mathcal{C},\cdot)$ and $I(X;Y) := I(\mathcal{A}_h,\cdot)$, $I(Y;Z) := I(\mathcal{B}_h,\cdot)$, $I(X;Z) := I(\mathcal{C}_h,\cdot)$. The subscript h indicates the h^{th} sub-channel representing a DMC.

Sketch of Proof for Theorem 6. We start with Equation (9) instead of Equation (14). The DPI is valid per sub-channel. So, for all h: $I(\mathsf{C}_h,\cdot) \leq \min[I(\mathsf{A}_h,\cdot), I(\mathsf{B}_h,\cdot)]$. As per Equation (4) we multiply both sides by $p(\zeta_h^-)$—the probability that the h^{th} channel is selected—and sum over h. This results in a DPI for transfer entropy,

$$TE(\mathcal{C},\cdot) \leq \min[TE(\bar{\mathcal{A}},\cdot), TE(\mathcal{B},\cdot)]. \tag{A3}$$

The tensor $\bar{\mathsf{A}}_h$ is itself the result of two cascaded channels represented by A_g and a tensor with elements p^g_{ih}. For these two DMC's the DPI is also valid, leading to:

$$\forall_{g,h}:\ I(\bar{\mathsf{A}}_h,\cdot) \leq I(\mathsf{A}_g,\cdot).$$

We now multiply both sides of this equation by $p(\zeta_h^-)p(\psi_g^-)$, and sum over h and g, resulting in $TE(\bar{\mathsf{A}},\cdot) \leq TE(\mathsf{A},\cdot)$. We can now rewrite Equation (A3) as

$$TE(\mathsf{C},\cdot) \leq \min[TE(\mathsf{A},\cdot), TE(\mathsf{B},\cdot)]. \tag{A4}$$

This implies that

$$TE(\mathcal{B}\odot\mathcal{A},\cdot) \leq \min[TE(\mathcal{A},\cdot), TE(\mathcal{B},\cdot)]. \tag{A5}$$

□

Appendix A.5. Theorem 7

Sketch of the proof of Theorem 7. If both $\mathcal{B} = \mathcal{C}\odot\mathcal{A}^\ddagger$ and $\mathcal{C} = \mathcal{B}\odot\mathcal{A}$ are valid, causal tensors can not distinguish a fork from a chain. There are two cases that need to be considered. In the first case, conditions are derived using the causal tensor relations. In the second case, we show that the pmfs impose a certain condition.

We start by combining $\mathcal{B} = \mathcal{C}\odot\mathcal{A}^\ddagger$ and $\mathcal{C} = \mathcal{B}\odot\mathcal{A}$:

$$\mathcal{B} = \mathcal{B}\odot\mathcal{A}\odot\mathcal{A}^\ddagger, \tag{A6a}$$
$$\mathcal{C} = \mathcal{C}\odot\mathcal{A}^\ddagger\odot\mathcal{A}. \tag{A6b}$$

These equations are valid when $\mathcal{B}\odot\mathcal{I}_1 = \mathcal{B}\odot\mathcal{A}\odot\mathcal{A}^\ddagger$ and $\mathcal{C}\odot\mathcal{I}_2 = \mathcal{C}\odot\mathcal{A}^\ddagger\odot\mathcal{A}$, with $\mathcal{I}_1$ and $\mathcal{I}_2$ identity causal tensors. Per definition identity tensors are noiseless. Because the causal tensors are stochastic tensors, their elements are nonnegative. The product of two stochastic tensors can only equal a noiseless tensor if and only if both $\bar{\mathsf{A}}$ and $\bar{\mathsf{A}}^\ddagger$ are noiseless. Along the same line of reasoning, we finally conclude that A and $\mathsf{A}^\ddagger$ are noiseless causal tensors because the averaging operation is in fact a matrix multiplication of two tensors.

The second case in which we cannot distinguish a fork and a chain follows from the pmf transformations:

$$\mathcal{B}\,|y\rangle = \mathcal{C}\odot\mathcal{A}^\ddagger\,|y\rangle\,, \tag{A7a}$$
$$\mathcal{C}\,|x\rangle = \mathcal{B}\odot\mathcal{A}\,|x\rangle\,. \tag{A7b}$$

The output from both the left-hand side and the right-hand side of these equations are probability mass functions. If they are indistinguishable, we cannot differentiate between a fork and a chain either.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2020 doi:10.20944/preprints202001.0106.v1

If both $\mathcal{B}$ and $\mathcal{C}$ are (almost) perfectly noisy causal tensors, both equations in Equation (A7) reduce to $|u\rangle = |u\rangle$, with $|u\rangle$ representing the uniform pmf. We cannot distinguish a chain from a fork. □

Appendix A.6. Theorem 8

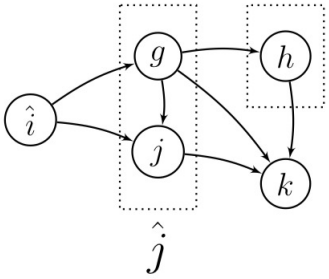

Figure A1. The graph related a chain of two causal channels. Both dotted boxes block the paths between $\hat{i}$ and k. The dotted box comprising the variables $\{g, j\}$ also blocks all paths between $\hat{i}$ and k. This box represents the variable $\hat{j}$.

The proof for Theorem 8 uses Figure A1. We furthermore do not use the Einstein summation convention.

Sketch of the proof of Theorem 8. According to the Law of Total Probability, $C^k_{\hat{i}} = \sum_h p^h_{\hat{i}} C^k_{h\hat{i}}$. Multiplying both sides of Equation (9) by $p^h_{\hat{i}}$, gives us

$$C^k_{\hat{i}} = \sum_{\hat{j}} \sum_h p^h_{\hat{i}} \bar{A}^{\hat{j}}_{h\hat{i}} B^k_{h\hat{j}}. \tag{A8}$$

We now express all stochastic tensors with the letter p instead of an A or a B. Using the fact that $p^h_{\hat{i}} p^{\hat{j}}_{h\hat{i}} = p^{\hat{j}h}_{\hat{i}}$, and $p^k_{h\hat{j}} = p^{kh}_{\hat{j}} / p^h_{\hat{j}}$, Equation (A8) can be written as:

$$C^k_{\hat{i}} = \sum_{\hat{j}} \sum_h p^{\hat{j}h}_{\hat{i}} p^{kh}_{\hat{j}} / p^h_{\hat{j}}. \tag{A9}$$

Because $p^{\hat{j}h}_{\hat{i}} = p^{\hat{j}}_{\hat{i}} p^h_{\hat{i}ij}$ and $p^{kh}_{\hat{j}} = p^k_{\hat{j}} p^h_{\hat{j}k}$ we get

$$C^k_{\hat{i}} = \sum_{\hat{j}} p^{\hat{j}}_{\hat{i}} p^k_{\hat{j}} \sum_h p^h_{\hat{i}ij} p^h_{\hat{j}k} / p^h_{\hat{j}}. \tag{A10}$$

Using $p^h_{\hat{i}ij} = p^{h\hat{i}}_{\hat{j}} / p^{\hat{i}}_{\hat{j}}$ and $p^h_{\hat{j}k} = p^{hk}_{\hat{j}} / p^k_{\hat{j}}$ we rewrite Equation (A10) as

$$C^k_{\hat{i}} = \sum_{\hat{j}} p^{\hat{j}}_{\hat{i}} p^k_{\hat{j}} \sum_h \frac{p^{h\hat{i}}_{\hat{j}}}{p^{\hat{i}}_{\hat{j}}} \frac{p^{hk}_{\hat{j}}}{p^k_{\hat{j}}} \frac{1}{p^h_{\hat{j}}}. \tag{A11}$$

Using a similar step as in the previous cases, Equation (A11) can be rewritten as

$$C^k_{\hat{i}} = \sum_{\hat{j}} p^{\hat{j}}_{\hat{i}} p^k_{\hat{j}} \sum_h \frac{p^{\hat{i}}_{h\hat{j}}}{p^{\hat{i}}_{\hat{j}}} \frac{p^k_{h\hat{j}}}{p^k_{\hat{j}}} p^h_{\hat{j}}. \tag{A12}$$

Applying the Causal Markov Condition on the graph in Figure A1, gives us $\hat{i} \perp\!\!\!\perp k | \{h, \hat{j}\}$, i.e., $p^{\hat{i}}_{h\hat{j}} p^k_{h\hat{j}} p^h_{\hat{j}} = p^{h\hat{i}k}_{\hat{j}}$. Because we sum over h, we sum out this variable as per Law of Total Probability. Using $A^{\hat{j}}_{\hat{i}} = p^{\hat{j}}_{\hat{i}}$ and $B^k_{\hat{j}} = p^k_{\hat{j}}$ we finally get the following expression

$$C_{\hat{i}}^{k} = \sum_{\hat{j}} A_{\hat{i}}^{\hat{j}} B_{\hat{j}}^{k} \frac{p_{\hat{j}}^{\hat{i}k}}{p_{\hat{j}}^{\hat{i}} p_{\hat{j}}^{k}}. \tag{A13}$$

So, if $\hat{i}$ and k are independent given $\hat{j}$, the theorem has been proven. When we apply the Causal Markov Condition to the graph in Figure A1, we see that $\hat{i} \perp\!\!\!\perp k|\hat{j}$. □

References

1. Guo, R.; Cheng, L.; Li, J.; Hahn, P.; Liu, H. A Survey of Learning Causality with Data: Problems and Methods **2018**.
2. Eichler, M. Causal inference with multiple time series: Principles and problems. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **2013**, *371*, 20110613. doi:10.1098/rsta.2011.0613.
3. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2009.
4. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438.
5. Vastano, J.A.; Swinney, H.L. Information transport in spatiotemporal systems. *Phys. Rev. Lett.* **1988**, *60*, 1773–1776. doi:10.1103/PhysRevLett.60.1773.
6. Dagum, P.; Galper, A.; Horvitz, E. Dynamic Network Models for Forecasting. Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1992; UAI'92, pp. 41–48.
7. Spirtes, P.; Glymour, C.; N., S.; Richard. *Causation, Prediction, and Search*; Mit Press: Cambridge, 2000.
8. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. doi:10.1103/PhysRevLett.85.461.
9. Lizier, J.T.; Prokopenko, M. Differentiating information transfer and causal effect. *The European Physical Journal B* **2010**, *73*, 605–615. doi:10.1140/epjb/e2010-00034-5.
10. Hyvärinen, A.; Zhang, K.; Shimizu, S.; Hoyer, P.O. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *J. Mach. Learn. Res.* **2010**, *11*, 1709–1731.
11. Runge, J.; Heitzig, J.; Petoukhov, V.; Kurths, J. Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy. *Phys. Rev. Lett.* **2012**, *108*, 258701. doi:10.1103/PhysRevLett.108.258701.
12. Duan, P.; Yang, F.; Chen, T.; Shah, S. Direct Causality Detection via the Transfer Entropy Approach. *Control Systems Technology, IEEE Transactions on* **2013**, *21*, 2052–2066. doi:10.1109/TCST.2012.2233476.
13. Sun, J.; Taylor, D.; Bollt, E. Causal Network Inference by Optimal Causation Entropy. *SIAM Journal on Applied Dynamical Systems* **2014**, *14*. doi:10.1137/140956166.
14. Turing, A.M. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **1937**, *s2-42*, 230–265, [http://oup.prod.sis.lan/plms/article-pdf/s2-42/1/230/4317544/s2-42-1-230.pdf]. doi:10.1112/plms/s2-42.1.230.
15. Copeland, B.J. The Church-Turing Thesis. In *The Stanford Encyclopedia of Philosophy*, Spring 2019 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2019.
16. Pearl, J. Causal Inference in Statistics: An Overview. *Statistics Surveys* **2009**, *3*, 96–146. doi:10.1214/09-SS057.
17. Eberhardt, F.; Scheines, R. Interventions and Causal Inference. *Philos. Sci.* **2007**, *74*. doi:10.1086/525638.
18. Ding, M.; Chen, Y.; Bressler, S.L. Granger Causality: Basic Theory and Application to Neuroscience, 2006.
19. Ghassami, A.; Kiyavash, N. Interaction information for causal inference: The case of directed triangle. 2017 IEEE International Symposium on Information Theory (ISIT), 2017, pp. 1326–1330. doi:10.1109/ISIT.2017.8006744.
20. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 1991.
21. Margolin, A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla-Favera, R.; Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC bioinformatics* **2006**, *7 Suppl 1*, S7. doi:10.1186/1471-2105-7-S1-S7.
22. Papoulis, A.; Pillai, S.U. *Probability, Random Variables, and Stochastic Processes*, fourth ed.; McGraw Hill, 2002.

23. 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.
24. Ahmed, S.S.; Roy, S.; Kalita, J.K. Assessing the Effectiveness of Causality Inference Methods for Gene Regulatory Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2018**, pp. 1–1. doi:10.1109/TCBB.2018.2853728.
25. Rashidi, B.; Singh, D.S.; Zhao, Q. Data-driven root-cause fault diagnosis for multivariate non-linear processes. *Control Engineering Practice* **2018**, *70*, 134 – 147. doi:https://doi.org/10.1016/j.conengprac.2017.09.021.
26. Dullemond, K.; Peeters, K. Introduction to Tensor Calculus. 2010.
27. Muroga, S. On the Capacity of a Discrete Channel. *Journal of the Physical Society of Japan* **1953**, *8*, 484–494. doi:10.1143/JPSJ.8.484.
28. Beckenbach, E.F. *Modern mathematics for the engineer: second series*; New York : McGraw-Hill, 1961.
29. Wibral, M.; Pampu, N.; Priesemann, V.; Siebenhühner, F.; Seiwert, H.; Lindner, M.; Lizier, J.T.; Vicente, R. Measuring information-transfer delays. *PloS one* **2013**.
30. Dean, T. *Network+ Guide to Networks*, 6th ed.; Course Technology Press: Boston, MA, United States, 2012.
31. Berge, C. *Graphs and Hypergraphs*; Elsevier Science Ltd.: Oxford, UK, UK, 1985.
32. Dirac, P.A.M. A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society* **1939**, *35*, 416–418. doi:10.1017/S0305004100021162.
33. Nauta, M.; Bucur, D.; Seifert, C. Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning and Knowledge Extraction* **2019**, *1*, 312–340. doi:10.3390/make1010019.
34. Wibral, M.; Wollstadt, P.; Meyer, U.; Pampu, N.; Priesemann, V.; Vicente, R. Revisiting Wiener's principle of causality — interaction-delay reconstruction using transfer entropy and multivariate analysis on delay-weighted graphs. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 3676–3679. doi:10.1109/EMBC.2012.6346764.
35. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* **1972**, *18*, 460–473. doi:10.1109/TIT.1972.1054855.
36. Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. *preprint* **2010**, *1004*.