Towards a framework for observational causality from time series: when Shannon meets Turing

David Sigtermans
ASML Netherlands B.V.
(Dated: November 11, 2019)

A tensor based formalism is proposed for inferring causal structures. This formalism enables us to determine the directionality of relations within a complex network. It furthermore allows us to differentiate between direct and indirect associations in the case of noisy data. Using this framework a Data Processing Inequality is proved to exist for Transfer Entropy. Once a causal graph has been inferred, the formalism enables simulating the behavior of the network.

I. INTRODUCTION

Exact knowledge about the functional relationships that fully determine the behavior of complex systems is a holy grail in the (applied) sciences and engineering. Several methods have been developed to arrive at causal or associational descriptions. The main difference between a causal and an associational description is that for a causal description experimentation is required, whereas for associational description (statistical) data requirements suffice [1]. Because interventions are not always possible, we have to make do with the data, a plethora of methods to infer causal structures from observational data have been developed, see for example [2–6]. None of these methods seem currently capable of both differentiating between direct and indirect associations (i.e. association via one or more mediators) and determining the directionality, within their own formalism.

In this paper a novel approach inspired by Turing machines [7] is proposed. If causal relations can be computed given the data, a Turing machine exists that “computes” causality, i.e., the causal relation is encoded in the transition function. Transfer Entropy [4] is a measure that can capture causal relations as far as encoded in the probability density functions [8]. Instead of inferring the transition functions of the related Turing machines, we derived a tensor formalism utilizing concepts from Information Theory [9]. This formalism: (1) is able to determine the directionality of relations within a complex network, (2) can differentiate between direct and indirect associations, and (3) enables simulating the behavior of the network using the inferred relations. We furthermore show that noise is needed for proper causal inference using our framework.

A. Outline

We start this paper with an introduction of aspects from Information Theory that are needed to derive our framework. Next (bivariate) Transfer Entropy (TE) is introduced. Transfer Entropy is capable of detecting directionality and cycles. Using concepts from Information Theory it is shown that TE allows for a tensor based formalism which gives rise to a specific set of calculation rules. We then show that this framework lets us differentiate between direct and indirect relationships. It is also used to derive conditions when this is not possible. We end this paper with an example to illustrate that we are indeed capable of detecting nonlinear relationships.

II. PRELIMINARIES

Statistical independence is foundational to causal inference [5], and therefore also to this paper. We will give a short overview of 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 exact same independence relations that follow from the underlying probability distributions. E.g. the faithfulness assumption for the chain \( X \rightarrow Y \rightarrow Z \) implies that \( X \) and \( Z \) are independent given \( Y \): \( p(X, Z|Y) = p(X|Y)p(Z|Y) \). This is denoted as \( X \perp\!\!\!\perp Z|Y \).

The Causal Markov Condition states that a process is independent of its non-effects, given its direct causes, i.e., its parents. This is relevant in the context of time series. We illustrate the Causal Markov Condition with an example that will be used later in this paper.

Example 1. Let \( i \) and \( g \) be the parents of \( j \), let \( g \) also be the parent of \( h \) and let \( i \) and \( j \) be non-effects of \( h \). According to the Causal Markov Condition \( i \) and \( h \) are independent given \( g \), i.e. \( p(j|g, h, i) = p(j|g, i) \).

III. INFORMATION THEORY

Information Theory was introduced in 1948 by C. Shannon [9]. It models association between random variables as resulting from a communication process between a sender and a receiver, i.e. transmission of a message from sender to receiver. A message consists of indexed realizations of random variables representing stationary ergodic processes. An input message is first encoded. In this paper we simplify encoding to describing the message with a finite alphabet. Each random variable has its own finite alphabet. The random variable \( X \) selects symbols from the alphabet \( \mathcal{X} \), the random variable \( Y \)
selects symbols from \( \mathcal{Y} \), and the random variable \( Z \) selects symbols from \( \mathcal{Z} \). Where \( \mathcal{X} = \{ \chi_1, \chi_2, \ldots, \chi_{|X|} \} \), \( \mathcal{Y} = \{ \psi_1, \psi_2, \ldots, \psi_{|Y|} \} \), and \( \mathcal{Z} = \{ \zeta_1, \zeta_2, \ldots, \zeta_{|Z|} \} \). The number of elements in the alphabet, i.e. 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 input symbol is send and a specific output symbol is received only depends on the alphabet symbol that was send. This implies that the communication process transforms the input probability mass function (pmf) into the output pmf. With \( x \in \mathcal{X} \) a realization of \( X \) and \( y \in \mathcal{Y} \) a realization of \( Y \) we have \( p(y) := \Pr\{Y = y\} \) and \( p(x) := \Pr\{X = x\} \) respectively.

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

### A. Mutual Information

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

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

the mutual information (MI), is nonnegative and symmetric in \( X \) and \( Y \). It represents the reduction in uncertainty about the random variable \( Y \) given that we have knowledge about the random variable \( X \).

It is intuitively clear that, given the information content of source data, in subsequent transmission steps the information can never increase. This is formalized in the Data Processing Inequality or DPI: processing of data can never increase the amount of information \( [10] \). For the cascade \( X \to Y \to Z \) 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.

### B. 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 (left hand side) and an output side (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 it’s own input vertex. Likewise, every output alphabet symbol has it’s own output vertex.

The simplest type of channel is the noisy discrete memoryless communication channel (DMC). 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 a consequence of the Law of Total Probability [11] because

\[
\Pr\{Y = \psi_j\} = \sum_i \Pr\{X = \chi_i\} \Pr\{Y = \psi_j|X = \chi_i\},
\]

with \( \Pr\{Y = \psi_j\} \) the \( j \)th element of \( p(y) \), and \( \Pr\{X = \chi_i\} \) the \( i \)th element of \( p(x) \). The transmission of data over a DMC transforms the probability mass function 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 [10].

Assuming a fixed (e.g. lexicographic) order of the alphabet elements, we can introduce an index notation for the pmf’s, e.g, \( p^i := \Pr\{Y = \psi_j\} \) and \( p^i := \Pr\{X = \chi_i\} \). In this paper every index is associated with a specific random variable. In Table I an overview is given.

| Process | Variable | Alphabet element | Index (input) | Index (past) | Index (output) |
|---------|----------|------------------|---------------|--------------|---------------|
| X       | x        | \( \chi \)       | \( i \)       | fc           | i             |
| Y       | y        | \( \psi \)       | \( j \)       | fc           | j             |
| Z       | z        | \( \zeta \)      | \( k \)       | fc           | k             |

### C. 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. This summation convention simplifies equations by implying summation over indices that appear once as an upper, or contra-variant index and once as a lower, or covariant index. Using these we rewrite Eq. (3) as

\[
p^i = p^i A^i_j.
\]

The covariant indices indicate the variables that we condition on. The row stochastic probability transition matrix elements represent the elements of the probability
transition tensor $\mathbf{A}$ \cite{14}. Using the standard notation i.o. the Einstein summation convention, MI can be rewritten as

$$ I(X;Y) = \sum_{i,j} p^j A^i_j \log_2 \frac{A^i_j}{\sum_{i'} p^{i'} A^{i'}_j}. \quad (5) $$

Mutual information 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 from 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.** Assume that $A^i_j = \delta^i_j$, i.e., the symbol received equals the symbol send. In this case Eq. (3) reduces to $I(X;Y) = \sum_i p^i \log_2 \frac{p^i}{\mathbb{E}[p^i]}$. 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 $\varepsilon$. Taking the limit $\varepsilon \to 0$ results in a mutual information $\to 0$. In other words, although there might be a perfect, i.e., noiseless, channel that represents 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, the probability transition tensors or measures based on elements of probability transition tensor should be used.

The earlier mentioned channel capacity is such a measure. It only depends on the elements of the probability transition tensor \cite{13}, e.g., $C_{XY} := \Gamma(A)$. In our example of perfect transmission with an arbitrarily small MI, the channel capacity only depends on the number of alphabet elements: $C_{XY} = \min \{ \log_2(|\mathcal{X}|), \log_2(|\mathcal{Y}|) \}$. This gives rise to the following definition

**Definition 1** (Normalized channel capacity). The normalized channel capacity is defined as

$$ \gamma(A) = \frac{\Gamma(A)}{\min \{ \log_2(|\mathcal{X}|), \log_2(|\mathcal{Y}|) \}} $$

Because the channel capacity is the maximal achievable mutual information for a specific channel, the earlier mentioned DPI is also applicable to the channel capacity.

**Corollary 1** (DPI for channel capacity). For the chain $X \to Y \to Z$ the DPI immediately implies that

$$ \Gamma(C) \leq \min \{ \Gamma(A), \Gamma(B) \}. \quad (6) $$

With $A$ representing the tensor of the transmission $X \to Y$, $B : Y \to Z$, and $C : X \to Z$.

The proof is straightforward and therefor omitted.

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. In its formulation it can be applied to all cases where observational data are available.

### IV. TRANSFER ENTROPY

Schreiber introduced Transfer Entropy in 2000 \cite{4}. Like MI it is non-parametric, but unlike MI it is an essentially asymmetric measure and as such it enables the differentiation between a source and a destination. It is an information theoretical implementation of Wiener’s principle of Causality \cite{15}: a cause combined with the past of the effect predicts the effect better than that the effect predicts itself. In contrast to Granger causality \cite{2}, Transfer Entropy is capable of capturing nonlinear relationships.

In this paper we use a slightly modified version which was shown to fully comply to Wiener’s principle of Causality by Wirral et al. It was proved that this modified TE is maximal for the real interaction delay \cite{13}. 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 $y^- := (y_{t-1}, \ldots, y_{t-\ell})$. The destination also depends on the source data $X$. With $\tau$ the finite interaction delay, it is assumed that for the input symbol $x^- \in (x_{t-\tau}, \ldots, x_{t-\tau-m})$, with $m \geq 0$. The alphabet for the past of $Y$ is $\mathcal{Y}^\ell$. The alphabet for the input is $\mathcal{X}^m$.

$$ T E_{X \to Y} = \sum_{x^- \in \mathcal{X}^m, y \in \mathcal{Y}^\ell} p(x^-, y, y-) \log_2 \frac{p(y|x^-, y^{-})}{p(y|y^-)}. \quad (7) $$

To be able to differentiate a cause from an effect, two hypotheses have to be assessed: (1) $X$ is the cause and $Y$ is the effect, and (2) $Y$ is the cause and $X$ is the effect. Per case the interaction delay that maximizes the respective TE is determined. If the resulting TE equals 0, it is assumed that there is no relation. Assuming that 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 with the largest TE is determined. (2) The optimal interaction delays are different: we assume that the hypothesis with the largest TE is determined. With loss of generality we assume in this paper that there are no cycles and that the interaction delays are all equal to 0.

Transfer Entropy is a conditional mutual information \cite{4}. It is therefore likely that it can be associated with communication channels. We start with conditioning the
MI from Eq. (1) on the event $y^- = \psi_g^-$ resulting in

$$ I(X; Y|\psi_g^-) = \sum_{x^- \in X^m \atop y \in Y} p(x^-, y|\psi_g^-) \log_2 \left[ \frac{p(y|x^-, \psi_g^-)}{p(y|\psi_g^-)} \right]. \quad (8) $$

Because $x^-$ and $y^-$ are the only parents of the output $y$, it follows from the Causal Markov Condition that the associated channel is memoryless. This sub-channel information quantifies the amount of information that is transmitted over the $g^{th}$ sub-channel. Transfer Entropy of Eq. (9) can now be expressed as

$$ TE_{X \to Y} = \sum_{\psi_g^- \in Y^d} p(\psi_g^-) I(X; Y|\psi_g^-). \quad (9) $$

### A. The causal channel

Equation (9) gives rise to a very specific communication channel: a channel with the topology of an inverse multiplexer. An inverse multiplexer consists of 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 (multiplexes) several input data streams into a single output data stream $\hat{y}$. 

**Definition 2 (Causal channel).** A causal channel is an inverse multiplexer in which the demultiplexer selects the sub-channel over which the data are sent 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 DMC. 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 1.

**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 $\sum_{\psi_g^- \in Y^d} p(\psi_g^-) I(X; Y|\psi_g^-)$, which is the definition of TE in Eq. (9).

### B. Tensor representation of a causal channel

Because every sub-channel of the causal channel represents a DMC, a causal channel is 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

$$ p^i_g = p^i_g A^i_{gi}. \quad (10) $$

The elements of the tensor $A$ are given by $A^i_{gi} = p(\psi_i|\chi^-_i, \psi_g^-)$. TE can now be rewritten as

$$ TE_{X \to Y} = \sum_{g, l, j} p^i_g p^j_l A^j_{gi} \log_2 \left[ \frac{A^j_{gi}}{\sum_{i'} p^i_{g'i} A^j_{gi}} \right]. \quad (11) $$

In a similar fashion as MI, it can be shown 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 calculation of the channel capacity for a causal channel is not trivial. We assume however that it is possible to determine the channel capacity.

When causal tensors are used to infer the underlying structure it opens the possibility of simulation once the structure has been determined, each edge in a directed causal graph has a corresponding causal tensor. As indicated in the introduction, the approach in the paper was inspired by a Turing machine. 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 pmf. To warrant the use of the adjective “causal” however, we have to show that within the framework of causal tensors we are capable to differentiate between direct and indirect associations. That this seems possible can be intuited when considering the chain $X \to Y \to Z$ (see Figure 2a). The relation $X \to Z$ is a resultant of the other relations, i.e. an indirect association. Within the framework of causal tensors we would expect that this resulting relation can be expressed in terms of the tensors of the other relations once the algebraic rules for manipulating the tensors are known.
V. CALCULATION RULES FOR CAUSAL TENSORS

Every sub-channel described by the causal tensor is a (row) stochastic tensor. Operations performed on these tensors should result in either scalars, stochastic vectors or stochastic tensors. The basic algebraic rules are well known because we can borrow them from linear algebra.

Without loss of generality we assume that a bivariate measurement of the relations within a system consisting of three random variables, results in the directed triangle being measured.

\[
\begin{align*}
B_{hi}^k &= p_{hi}^g A_{gi}^j B_{kj}^h, \\
\end{align*}
\]  

(13)

For readability we moved the proof to the appendix. If both A and B represent DMC’s we get the simpler, well known product rule for a chain of DMC’s.

**Corollary 2** (Product rule for a chain consisting 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

\[
C_{hi}^k = A_{hi}^j B_{kj}^h.
\]

(14)

Let A be the causal tensors of two DMC’s and the (\(\hat{\cdot}\)) operation because it is indicative of the origin of these conditional probabilities. We can now rewrite Eq. (13) as

\[
C_{hi}^k = \hat{A}_{hi}^j B_{kj}^h.
\]

(15)

The causal tensor \(\hat{A}_{hi}^j\) is a trivariate tensor: index \(\hat{i}\) is associated with random variable X, index \(\hat{j}\) is associated with random variable Y, and index \(\hat{h}\) is associated with random variable Z respectively. According to corollary 2, this equation can be interpreted as representing two DMC’s in series. This means that we have an alternative structure for two causal channels in series as depicted in Figure 1. Because the Data Processing Inequality is applicable to a cascade of DMC’s the alternative structure suggests that there is a DPI for Transfer Entropy.

A. The chain structure

First let the chain \(X \rightarrow Y \rightarrow Z\) be the ground truth. Additional to Eq. (11), \(p_{ij}^k = p_{ij}^g A_{ji}^j\), there are two other causal tensors: \(B : Y \rightarrow Z\), and \(C : X \rightarrow Z\). Because it is a straightforward exercise we leave it to the reader to confirm that

\[
\begin{align*}
p_{hi}^k &= p_{hi}^g B_{hj}^k, \\
p_{hi}^k &= p_{hi}^g C_{hi}^k.
\end{align*}
\]

(12a)

The index \(\hat{i}\) in Eq. (12a) is the index related to the cause \(\hat{x} \in \mathcal{X}'\) of Z. The index \(\hat{i}\) in Eq. (11) is the index related to the cause \(\hat{x} \in \mathcal{X}'\) of Y. The Markov property immediately implies that in both cases we can use the same cause vector, indicated by say \(\hat{i}\), as long as \(m \geq m'\).

**Theorem 2** (Product rule for a chain). Let A and B be the causal tensors of two causal channels 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

\[
C_{hi}^k = p_{hi}^g A_{gi}^j B_{kj}^h.
\]

(13)

B. The fork structure

Assume that the fork is the ground truth. The goal is to express the indirect association represented by B in terms of the other causal tensors. First of all we notice that the input distribution can be reconstructed from the output distribution.

**Definition 3** (Reconstruction). The \(\dagger\)-operation, or reconstruction operation, reconstructs the source distribution, conditioned of the past of the destination, from the destination distribution, conditioned of the past of the destination:

\[
p_{ij}^g = p_{ij}^g A_{ji}^{\dagger},
\]

(16)

with \(A_{ji}^{\dagger} = p_{ji}^g\).

The \(\dagger\)-operation applied to the directed graph \(X \rightarrow Y\), results in the graph \(X \leftarrow Y\). This implies that in the framework of causal tensors, a fork has equivalent chains.
Theorem 3 (Fork-chain equivalence). The fork $X \rightarrow Y + X \rightarrow Z$ is equivalent to the chain $Y \rightarrow X \rightarrow Z$ and to the chain $Y \leftarrow X \leftarrow Z$.

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).

$$B^k_{hj} = A^i_{hj} C^k_{hi}, \quad \text{with } A^i_{hj} := p^i_{hj} A^i_{gj},$$

$$B^j_{gk} = C^i_{gk} A^j_{si}, \quad \text{with } C^i_{gk} := p^i_{gk} C^i_{hk}.$$  \hspace{1cm} (17a)

Equation (17a) is applicable in the case depicted in Figure 2a, i.e. a bivariate measurement between $Y$ and $Z$ results in an indirect association with $Y$ as the cause and $Z$ as the effect. The equivalent chain is $Y \rightarrow X \rightarrow Z$. In the case that for the indirect association $Z$ is the cause and $Y$ the effect, Eq. (17b) is applicable. The equivalent chain in that case is $Y \leftarrow X \leftarrow Z$ is.

C. The v-structure & 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 2a. However, investigating structures with a collider, the v-structure and the more general directed triangle, will result in the important concept of interaction. So, lets assume that the ground truth is the directed triangle. We now have to introduce the multivariate relation $D: \{X, Y\} \rightarrow Z$. This relation leads to the additional linear transformation

$$\hat{p}_h^k = \hat{p}_h^j D^k_{hij},$$

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

Lemma 1 (Causal Tensor Contraction). In the case of a directed triangle we can express the causal tensors in terms of the interaction tensor;

$$B^k_{hj} = A^i_{hj} D^k_{hij},$$

$$C^k_{hi} = A^j_{hi} D^k_{hij}. \hspace{1cm} (19a)$$

We will only derive this relation for Eq. 19a.

Sketch of Proof. First we note that $\hat{p}_h^j = \delta_j^j p^i_{hj} p^i_{hj}$. With this Eq. 18 is rewritten as:

$$\hat{p}_h^k = \delta_j^j p^i_{hj} D^k_{hij}. \hspace{1cm} (20)$$

By changing the order of $\delta_j^j$ and $\hat{p}_h^k$ we get $\hat{p}_h^k = p^i_{hj} (\delta_j^j p^i_{hj} D^k_{hij})$. Combining this with Eq. 12a results in an expression for $B^k_{hj}$:

$$B^k_{hj} = \delta_j^j p^i_{hj} D^k_{hij}.$$  \hspace{1cm} (21)

Because $\delta_j^j p^i_{hj} = p^i_{hj}$ we get Eq. 19a.

From Eq. 19 it follows that $B$ and $C$ are the result of a cascade involving $A$, $A^j$ and $D$. The graphs represented by Figures 2b and 2c support the tensor relations, e.g., $X \rightarrow \{X, Y\} \rightarrow Z$ is equivalent to the cascade of the inverse multiplexers represented by $A$ and $D$ resulting in $C$. Figures 2d and 2e 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 [17]. 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. If one of the relations is indirect, no interaction takes place.

Theorem 4 (indirect causes do not contribute to an interaction). The interaction tensor only depends on the direct causes, not on indirect causes. So, if and only if the chain is the ground truth

$$D^k_{hij} = B^k_{hij},$$  \hspace{1cm} (20)

if and only if the fork is the ground truth

$$D^k_{hij} = C^k_{hi}. \hspace{1cm} (21)$$

For the proof we use the fact that the elements of a causal tensor are conditional probabilities. Again due to the fork-chain equivalence, we only need to proof it for a chain.

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

An immediate consequence of this theorem is that in general a fork, a chain and a directed triangle can be distinguished. The conditions under which it is not possible will be derived later.

Corollary 3. If and only if the chain is the ground truth

$$B^k_{hj} \neq A^i_{hj} C^k_{hi}.$$  \hspace{1cm} (22)

If and only if the fork is the ground truth

$$C^k_{hi} \neq A^i_{hi} B^k_{hj}. \hspace{1cm} (23)$$

In the case of a directed triangle, neither Eq. 13 nor Eq 17 are valid.
We will only proof this in the case of a chain.

**Sketch of Proof.** If the ground truth is a chain, the ground truth is not a fork. According to theorem 1, \( D_{hij}^k \neq C_{hi}^k \). Combining this with Eq. (10a) results in \( B_{hj}^k \neq \bar{A}_{hi}^k C_{hi}^k \).

In the following two examples we will illustrate that indirect associations do not interact. Without loss of generality we assume that the causal tensors represent DMC’s.

**Example 3.** Let the chain \( X \rightarrow Y \rightarrow Z \) be the ground truth. With \( A = (\frac{1}{2} \frac{1}{2} 0) \) and \( B = (\frac{1}{2} \frac{1}{2} \frac{1}{2}) \), the indirect association is represented by the causal tensor \( C = A \cdot B \Rightarrow C = (\frac{1}{3} \frac{1}{3} \frac{1}{3}) \). Assume that \( \bar{x} = (\frac{2}{3} \frac{2}{3}) \). The pmf for \( \bar{y} \) equals \( y = \bar{x}A \Rightarrow \bar{y} = (\frac{4}{9} \frac{1}{9}) \). Using the relation \( \hat{x} = yA^\dagger \) the reader can verify that \( A^\dagger = (\frac{1}{3} \frac{1}{3} \frac{1}{3}) \). Because

\[
\bar{A}_{hi}^i C_{hi}^k = (\frac{7}{12} \frac{17}{36} \frac{17}{36}) \Rightarrow B_{hj}^k \neq \bar{A}_{hi}^i C_{hi}^k, \text{ Eq. (23)} \text{ is indeed valid.}
\]

In the following example it is assumed that the fork is the ground truth.

**Example 4.** Let the fork \( X \rightarrow Y + X \rightarrow Z \) be the ground truth. Assume that \( A = (\frac{1}{2} \frac{1}{2} 0) \), \( C = (\frac{1}{2} \frac{1}{2} \frac{1}{2}) \) and \( \bar{x} = (\frac{2}{3} \frac{2}{3}) \). The pmf for \( \bar{y} \) equals \( \bar{y} = \bar{x}A \Rightarrow \bar{y} = (\frac{4}{9} \frac{1}{9}) \). As in the previous example \( A^\dagger = (\frac{1}{3} \frac{1}{3} \frac{1}{3}) \). The indirect association is represented by the causal tensor \( B = A^\dagger \cdot C \Rightarrow B = (\frac{7}{12} \frac{17}{36} \frac{17}{36}) \). Because \( \bar{A}_{hi}^i B_{hj}^k = (\frac{12}{18} \frac{27}{36} \frac{27}{36}) \Rightarrow C_{hi}^k \neq \bar{A}_{hi}^i B_{hj}^k, \text{ Eq. (22)} \text{ is valid.}

**D. Toward a causal tensor algebra**

The calculation rules for causal tensors follow from probability theory [11], Pearl’s theory of causality [1] and linear algebra. From the examples and derivations thus far we have seen that the operations on and with causal tensors follow very specific rules. These rules can be used to simplify notations even more.

The earlier introduced row stochastic causal tensors \( A, B \) and \( C \) are used with their respective indices. The stochastic row vectors, i.e. pmfs, are defined as \( \bar{x}, \bar{y} \) and \( \hat{x} \) so that: \( \bar{x}A = \bar{y}, \bar{y}B = \hat{x} \) and \( \hat{x}C = \bar{z} \). Furthermore the notation \( \{\cdot\} \) is used to indicated the elements of a tensor.

**Definition 4.** The channel averaging operator \( \{\cdot\} \) applied to a causal tensor \( A \) is defined as

\[
\forall_{h,i,j} \{ A \} := p^g_{hi} A_{gi}^j,
\]

with \( p^g_{hi} \), a trivariate row stochastic tensor.

The averaging operator plays a role in cascades.

**Definition 5.** The causal tensor cascading operator \( \odot \) applied to two causal tensors \( A \) and \( B \) is defined as a tensor contraction,

\[
A \odot B := A \cdot B, \text{ with } \forall_{h,j,k} \{ \bar{A} \} := \bar{A}_{hi}^j B_{hj}^k.
\]

The number of unique indices of the resulting tensor is always less than the total unique number of indices of the constituting tensors.

Because we use row stochastic tensors, a cascade is read from left to right, e.g., \( \bar{x}A \odot B \) is the transformation of the pmf \( \bar{x} \) via the operator \( A \). The resulting output pmf is then transformed via the operator \( B \).

**Definition 6.** The reconstruction operation \( \dagger \) reconstructs the input pmf from the output pmf

\[
\bar{x}A = \bar{y} \Leftrightarrow \bar{x} = \bar{y}A^\dagger.
\]

If \( \{A\} = A_{yi}^j \), then \( \{A^\dagger\} = A_{yi}^j \).

**Definition 7.** The identity causal tensor \( I \) is defined as

\[
\{I\} = \forall_k : I_{i,j,k} = \delta^j_i
\]

These definitions lead to the following properties (the proofs are straightforward and therefore omitted):

1. \( \{(\cdot)\} \) is distributive: \( \bar{A} \cdot B \cdot C = \bar{A} \cdot (B \cdot C) \)
2. \( \odot \) is associative: \( (A \odot B) \odot C = A \odot (B \odot C) \)
3. \( \odot \) is not commutative: \( A \odot B \neq B \odot A \)
4. \( \forall A \exists I : I \odot A = A \)

**VI. INFERRING STRUCTURES WITH CAUSAL TENSORS**

In this section we discuss some of the 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” (this will be defined later in this paper).
A. The 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 5 (Data processing inequality for TE). For the chain $X \to Y \to Z$ the following inequality holds

$$TE_{X \to Z} \leq \min \{TE_{X \to Y}, TE_{Y \to Z}\}. \quad (24)$$

For the proof a simplified notation for Transfer Entropy and mutual information is used. We write these measures as a function of pmfs, indicated by (·) and the respective tensor: $TE_{X \to Y} := TE(A, ·)$, $TE_{Y \to Z} := TE(B, ·)$, $TE_{X \to Z} := TE(C, ·)$ and $I(X; Y) := I(A, ·)$, $I(Y; Z) := I(B, ·)$, $I(X; Z) := I(C, ·)$. The subscript $h$ indicates the $h^{th}$ sub-channel representing a DMC.

Sketch of Proof. From Eq. (13), $C_h = \bar{A}_h \cdot B_h$, it follows that for a chain the DPI is valid per sub-channel.

$$\forall h : I(C_h, ·) \leq \min[I(\bar{A}_h, ·), I(B_h, ·)].$$

As per Eq. (9), multiplying both sides by $p(\zeta_h^i)$, i.e. the probability that the $h^{th}$ channel is selected, and summing over $h$, results in

$$TE(C, ·) \leq \min\{TE(\bar{A}, ·), TE(B, ·)\}, \quad (25)$$

a DPI for Transfer Entropy. The tensor $\bar{A}_h$ is itself the result of two cascaded channels represented by $A_g$ and a tensor with elements $p_i^{gh}$. For these two DMC’s the DPI is also valid, leading to:

$$\forall g, h : I(\bar{A}_h, ·) \leq I(A_g, ·).$$

We now multiply both sides of this equation by $p(\zeta_h^i)p(\zeta_g^i)$, and sum over $h$ and $g$, resulting in

$$TE(\bar{A}, ·) \leq TE(A, ·).$$

We can now rewrite Eq. (25) as

$$TE(C, ·) \leq \min\{TE(A, ·), TE(B, ·)\}, \quad (26)$$

or, equivalently,

$$TE(A \odot B, ·) \leq \min\{TE(A, ·), TE(B, ·)\}. \quad (27)$$

A similar DPI also exists for the channel capacity for causal tensors (see Eq. (20)).

B. Differentiating between direct and indirect associations with causal tensors

We have shown earlier that in general a fork, a chain and a directed triangle are distinguishable (see corollary 3). We now investigate in more detail under what conditions this is not possible.

Definition 8 (Perfect noisy relation). Iff all causal tensor elements are equal, the relation is a perfect noisy relation. The related causal tensor is called the perfect noisy causal tensor.

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 tensors is the noiseless causal tensor.

Definition 9 (Noiseless causal tensor). The elements of a noiseless causal tensor satisfy $\forall_{h,i,j} A_{hi} = 1$ and $\forall_{h,j} A_{hi} = 1$.

The reader can verify by using Eq. (11) that for any input pmf $TE = \log_2 \left[ \sum_i 1 \right]$. Because the channel capacity of a noiseless channel only depends on the number of alphabet elements, $C_{XY} = \min \{ \log_2(|X|), \log_2(|Y|) \}$, 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.

Theorem 6 (Perfectness). We are not able to differentiate between direct and indirect relations if: (1) all relations are perfectly noiseless, or (2) the relations are perfectly noisy.

Sketch of Proof. If both $B = \bar{A}^\dagger \cdot C$ and $C = \bar{A} \cdot B$ are valid, causal tensors can not distinguish a fork from a chain. There are two cases that need to 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 $B = \bar{A}^\dagger \cdot C$ and $C = \bar{A} \cdot B$:

$$B = \bar{A}^\dagger \cdot \bar{A} \cdot B, \quad (28a)$$

$$C = \bar{A} \cdot \bar{A}^\dagger \cdot C. \quad (28b)$$

These equations are valid when $I_1 = \bar{A}^\dagger \cdot \bar{A}$ and $I_2 = \bar{A} \cdot \bar{A}^\dagger$, with $I_1$ and $I_2$ 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 iff both $\bar{A}$ and $\bar{A}^\dagger$ are noiseless. Along the same line of reasoning we finally arrive to the conclusion that $\bar{A}$ and $\bar{A}^\dagger$ are noiseless causal tensors because the averaging operation is in fact a matrix multiplication of two tensors.

The second case in which a fork and a chain can not be distinguished follows from the pmf transformations:

$$\bar{y}B = \bar{y}\bar{A}^\dagger \cdot C, \quad (29a)$$

$$\bar{x}C = \bar{x}\bar{A} \cdot B. \quad (29b)$$
The output from both the left hand side and right hand side of these equations are probability mass functions. If they are indistinguishable, we can’t differentiate between a fork and a chain either. Assume that both B and C are perfect noisy causal tensors. With \( u(y) \) and \( u(x) \) representing the respective uniform pmfs, Eq. (29) reduces to

\[
\begin{align}
  u(y) &= u(y), \\
  u(x) &= u(x).
\end{align}
\]

\[\text{(30a) and (30b)}\]

In [18] an example is given of two perfect noisy relations that interacted resulting in a noiseless transmission. In other words, perfect noisy causal tensors can interact in such a way that the resulting interaction tensor is noiseless. On the other hand, perfect noiseless relations imply maximal redundancy within a data set.

C. Causal inference steps

To finalize the causal tensor framework as discussed so far, a short summary of the (implicitly) proposed steps is given. We assume that: (1) the data are time equidistant, (2) \( \ell \) 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) 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.
6. Use the product rule to determine if the indirect relations are indeed indirect.
7. Determine the interaction tensor for perfect noisy relations that collide.
8. If the network is used for simulation, determine the interaction tensors for all \( \psi \)-structures.

VII. EXPERIMENT

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

A. Ulam map

For the first experiment we use the one-dimensional lattice of unidirectional coupled maps \( x_{n+1}^m = f(x_{n}^m, 1) \). Information can only be transferred from \( x_{n}^{m-1} \) to \( x_{n}^m \). 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 [3]. We chose an alphabet consisting of 4 symbols. The quantization consisted of simple binning. Furthermore we chose \( \ell = m = 1 \) (see Eq. (7)). Instead of maximizing TE we maximized the channel capacity to determine the optimal delay. An approximation that satisfies the boundaries that follow from Eq. (30) was used,

\[
\hat{\Gamma}(A) = \sum_g p(\psi_g^y \Gamma(A_g)),
\]

To determine the channel capacities the Blahut-Arimoto algorithm was used [19]. The delays were varied between 1 and 20. The Channel capacity was maximal for a delay of 1 sample. As can be seen from Figure 3 causal tensors lead to a similar result as Transfer Entropy.

B. Coupled Ornstein-Uhlenbeck processes

In the second experiment we demonstrate our approach using a system of four coupled Ornstein-Uhlenbeck processes [32]:

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

with independent unit variance white noise processes \( \eta \). The integration time step was \( dt = 0.01s \) and the sampling interval \( \Delta s = 100s \). A binary encoding scheme was used. 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. On the other hand cascading very noisy channels reduces the probability of detecting an indirect relation. This is illustrated in Figure 3 no pruning was needed. This experiment shows that causal tensors are indeed capable of detecting the underlying structure.
delays that maximized the channel capacity are also shown.

Theorem 2 (Product rule for a chain) an edge is represented by a tensor i.o. a scalar.

simulating the behavior of the inferred system because proved to exist for TE. Finally, the formalism allows for Using this formalism a Data Processing Inequality was we showed the importance of noise within this formalism.

differentiate between direct and indirect associations and inferred. Theorems were established that allow us to tensor formalism with which causal structures can be

∀ Lemma 2.

Let the tensor

\[ \mathbf{A} \] represent the resulting indirect causal channel that must be measured in a bivariate approach.

FIG. 4. (a) The causal structure for the Ornstein-Uhlenbeck system of Eq.(32). 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 20

structures at different time series lengths. The confidence

VIII. CONCLUSION

To conclude, we used Transfer Entropy to come to a tensor formalism with which causal structures can be inferred. Theorems were established that allow us to differentiate between direct and indirect associations and we showed the importance of noise within this formalism. Using this formalism a Data Processing Inequality was proved to exist for TE. Finally, the formalism allows for simulating the behavior of the inferred system because an edge is represented by a tensor i.o. a scalar.

Appendix: Proof of product rule

Theorem 2 (Product rule for a chain). Let \( \mathbf{A} \) and \( \mathbf{B} \) be the causal tensors of two causal channels in series and let the tensor \( \mathbf{C} \) represent the resulting indirect causal channel that must be measured in a bivariate approach. The tensor elements of \( \mathbf{C} \) are given by

\[
C_{hi}^k = p_{ghi}^k \mathbf{A}_{gj}^i \mathbf{B}_{hi}^k.
\]  

(13)

For the proof we need to introduce two lemma’s.

Lemma 2. \( \forall g: \mathbf{B}_{ghi}^k = \mathbf{B}_{ghi}^k \).

Sketch of Proof. 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, e.g. \( g \), and the index related to the output, e.g. \( j \) appear in the same tensor we are allowed to replace the output index by the input index. In our example this means we are allowed to replace \( j \) by \( j \) as long as we ensure that \( \psi_j^\ast = \{ \psi_j, \psi_{j}^\prime \} \). This is always possible due to the Markov property: we either enlarge the cardinality of \( \psi_j^\ast \) or \( \psi_{j}^\prime \). \( \square \)

Lemma 3. For the chain \( X \rightarrow Y \rightarrow Z \) we have \( \mathbf{A}_{ghi}^j = \mathbf{A}_{ghi}^j \).

For the proof we refer to example [1]

Sketch of Proof. Because of the Law of Total Probability we are allowed to condition Eq.(10) on \( h \) and both Eq.(12a) and Eq.(12b) on \( g \). This leads to

\[
p_{ghi}^j = p_{ghi}^j \mathbf{A}_{ghi}^j;
\]  

(A.1a)

\[
p_{ghi}^k = p_{ghi}^k \mathbf{B}_{ghi}^k;
\]  

(A.1b)

\[
p_{ghi}^k = p_{ghi}^k \mathbf{C}_{ghi}^k;
\]  

(A.1c)

Substituting the expression for \( p_{ghi}^j \) of Eq.[A.1a] in Eq.[A.1b] and combining the result with Eq.[A.1c] gives us

\[
\mathbf{C}_{ghi}^k = \mathbf{A}_{ghi}^j \mathbf{B}_{ghi}^k.
\]  

(A.2)

Using lemma [2] and lemma [3] this can be rewritten as

\[
\mathbf{C}_{ghi}^k = \mathbf{A}_{ghi}^j \mathbf{B}_{ghi}^k.
\]  

(A.3)

Finally we multiply both sides with \( p_{ghi}^g \). As the reader can confirm, the term \( p_{ghi}^g \mathbf{C}_{ghi}^k \) equals \( \mathbf{C}_{ghi}^k \) leading to Eq.(2).

\( \square \)

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. Part of the work was supported by ASML PI System Diagnostics.
[1] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[2] Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. Granger causality: Basic theory and application to neuroscience, 2006.

[3] Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Phys. Rev. Lett.*, 108:258701, Jun 2012.

[4] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[5] Peter Spirtes, Clark Glymour, Scheines N., and Richard. *Causation, Prediction, and Search*. Mit Press: Cambridge, 2000.

[6] A. Ghassami and N. Kiyavash. Interaction information for causal inference: The case of directed triangle. In 2017 IEEE International Symposium on Information Theory (ISIT), pages 1326–1330, June 2017.

[7] A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 01 1937.

[8] J. T. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *The European Physical Journal B*, 73(4):605–615, Feb 2010.

[9] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.

[11] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, fourth edition, 2002.

[12] Kees Dillemond and Kasper Peeters. Introduction to tensor calculus. 2010.

[13] S. Muroga. On the Capacity of a Discrete Channel. *Journal of the Physical Society of Japan*, 8:484–494, July 1953.

[14] Edwin F Beckenbach. *Modern mathematics for the engineer: second series*. New York : McGraw-Hill, 1961.

[15] Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T. Lizier, and Raul Vicente. Measuring information-transfer delays. *PloS one*, 2013.

[16] Tamara Dean. *Network+ Guide to Networks*. Course Technology Press, Boston, MA, United States, 6th edition, 2012.

[17] C. Berge. *Graphs and Hypergraphs*. Elsevier Science Ltd., Oxford, UK, UK, 1985.

[18] Ryan G. James, Nix Barnett, and James P. Crutchfield. Information flows? a critique of transfer entropies. *Physical Review Letters*, 116(23), jun 2016.

[19] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, July 1972.