entropy

MDPI

Article

Information Theoretic Causal Effect Quantification

Aleksander Wieczorek * **and Volker Roth**

Department of Mathematics and Computer Science, University of Basel, CH-4051 Basel, Switzerland; volker.roth@unibas.ch

* Correspondence: aleksander.wieczorek@unibas.ch

Received: 31 August 2019; Accepted: 30 September 2019 ; Published: 5 October 2019

Abstract: Modelling causal relationships has become popular across various disciplines. Most common frameworks for causality are the Pearlian causal directed acyclic graphs (DAGs) and the Neyman-Rubin potential outcome framework. In this paper, we propose an information theoretic framework for causal effect quantification. To this end, we formulate a two step causal deduction procedure in the Pearl and Rubin frameworks and introduce its equivalent which uses information theoretic terms only. The first step of the procedure consists of ensuring no confounding or finding an adjustment set with directed information. In the second step, the causal effect is quantified. We subsequently unify previous definitions of directed information present in the literature and clarify the confusion surrounding them. We also motivate using chain graphs for directed information in time series and extend our approach to chain graphs. The proposed approach serves as a translation between causality modelling and information theory.

Keywords: directed information; conditional mutual information; directed mutual information; confounding; causal effect; back-door criterion; average treatment effect; potential outcomes; time series; chain graph

1. Introduction

Causality modelling has recently gained popularity in machine learning. Time series, graphical models, deep generative models and many others have been considered in the context of identifying causal relationships. One hopes that by understanding causal mechanisms governing the systems in question, better results in many application areas can be obtained, varying from biomedical [1,2], climate related [3] to information technology (IT) [4], financial [5] and economic [6,7] data. There has also been growing interest in using causal relationships to boost the performance of machine learning models [8].

1.1. Overview of Relevant Frameworks of Causality Modelling

Two main approaches to describing causality have been established. One is the Neyman-Rubin causal model or the potential outcomes framework. Its foundational idea is that of two counterfactual statements which are considered (e.g., application of a treatment or lack thereof) along with their effect on some variable of interest (e.g., recession of a disease). This effect (difference between the two counterfactual outcomes) is then measured with the average causal effect. Since for a single data point only one counterfactual is observed, any quantification in the Neyman-Rubin potential outcome model entails a fundamental missing data problem. The other main approach to modelling causal relationships are frameworks based on graphical models, predominantly directed acyclic graphs (DAGs). Such models are akin to Bayesian networks but are imbued with a causal interpretation with the idea of an intervention performed on a node. When a variable is intervened on, any influence of other nodes on it is suppressed and the distribution of the remaining variables is defined as their

interventional distribution. The effect of an intervention on a given node is then inferred given the structure of the entire DAG. Such DAGs along with inference rules concerning interventions (called causal calculus) were formalised by Pearl and are referred to as Pearlian graphs.

Regardless of the assumed framework, causal effects (understood as counterfactual outcomes or interventional distributions) can be directly estimated only in the presence of randomised experimental (also referred to as *interventional*) data, meaning that any counterfactual outcome or interventional distribution can be measured. Since this is infeasible in many application areas (e.g., effects of smoking or income cannot be obtained this way), attempts to quantify causal effects with observational data only have evolved into an active field within causal models.

Causal reasoning, within any of the assumed frameworks, can be formulated as one of two fundamental questions. Firstly, one can ask which variables influence which and in what order, that is, which counterfactual (or intervening on which nodes) will have an effect on a particular variable and which other variables have to be taken into account while measuring this effect. This is referred to as *causal induction*, *causal discovery* or *structure learning*, since it corresponds to learning the structure of the arrows of the Pearlian graph in the Pearlian framework. Secondly, once the structure of the causal connections between variables has been learnt or assumed, one can ask how to quantify the causal effect of one variable on another, for example, with the aforementioned average causal effect or the interventional distribution of the effect. This in turn is called *causal deduction*. The former question can be tackled with experimental data or exploiting conditional independence properties of the observable distribution (with algorithms such as PC [9] or IC [10]).

The answer to the latter question with observational data commonly involves a two-step procedure.

- First, a set of variables confounding the cause and effect variables is found. Confounding of variables X and Y by Z is a notion that formalises the idea of Z causally influencing (directly or not) both X and Y thus impeding the computation of the direct causal influence of X on Y. In a Pearlian graph, a set of confounders Z can be identified with rules of the causal calculus or with graphical criteria called the back-door criterion and front-door criterion. In the Neyman-Rubin potential outcome framework, a cognate criterion for a set Z is called strong ignorability and Z is frequently referred to as a set of sufficient covariates.
- After a set of confounders, or sufficient covariates, Z has been identified, the effect of X on Y is quantified. If such a Z exists, this can be done with only observational data. In the Pearlian setting, this amounts to the computation of the interventional distribution of Y given the intervention on X, which can be shown to be equal to conditioning on or adjusting for Z in the observational distribution of X, Y, Z. In the Neyman-Rubin potential outcome framework, the effect of X on Y is frequently measured with the average causal effect of X on Y, that is, the difference between expectations for the two potential outcomes. Even though exactly one potential outcome is observed, one can estimate the distribution of the missing one with observational data if one conditions on Z.

Thus, one first identifies the set of confounders and then uses them to draw conclusions about causal effects from observational data.

1.2. Causality Modelling and Information Theory

The goal of this paper is to formulate a comprehensive description of causal deduction as described above in terms of information theory. In particular, we relate to the most common frameworks of causality modelling sketched in Section 1.1 and provide a novel, rigorous translation of the most common causal deduction methods into the language of information theory.

Previous approaches to express causal concept with information theory were based on the notion of directed information. These approaches, however, were either limited to adjusting the concept of *Granger causality* to time series (which lead to a number of inconsistent definitions of directed information for time series) or only amounted to the first step of the causal deduction procedure.

Different approaches were also based on different definitions of directed information for time series and general DAGs. In the following, we clearly motivate *directed information* as the measure of *no confounding* and conditional *mutual information* as the measure that quantifies *causal effect* in a unconfounded setting (i.e., where confounding variables have been adjusted for). We also unify the definitions of directed information for general DAGs and time series and extend the definition to chain graphs which makes it possible to introduce a unique definition of directed information for time series. Finally, we respond to criticisms of directed information that were formulated by different authors as intuition-violating counterexamples and show how our approach allows for a comprehensive causal interpretation of directed information along with regular mutual information.

1.3. Related Work on Directed Information and Its History

Directed information has been defined differently by various authors. The main discrepancies stem from the level of definition generality (for two time series [11], multiple time series [12,13], generalised structures [14] and DAGs [15]) and from treatment of instantaneous time points in time series [13]. The former has led to a chain of definitions being generalisations of one another while the latter produced inconsistent definitions for time series. We propose an approach subsuming the different definitions: it is based on extending the most general definition to chain graphs in Section 3.

Directed information was originally introduced as a measure for feedback in discrete memoryless channels [11,16]. It was subsequently imbued with a causal interpretation by noting its similarity to the concept of Granger causality [17,18] between two time series: time series T_1 Granger-causes T_2 if the past of T_1 provides more information about the present of T_2 than the past of T_2 does. This resulted in two strategies for formalising directed information: one can assume the instantaneous time points of T_1 to be a part of the past and condition on them [11,19–21] or not [13,16,22,23]. Extensions to account for side information [19] and stochastic processes in continuous time [24] have also been put forward.

The original definition of directed information for discrete memoryless channels was subsequently extended to any time ordering with directed stochastic kernels [14]. This definition, in turn, was a special case of the definition introduced by Raginsky [15]: directed information as KL divergence of interventional and observational distributions. It was shown in the same paper that the conditional version of directed information being zero is equal to the backdoor criterion [25] for no confounding.

Comparing observational and interventional distributions in time series was also considered [12,26,27] and resulted in conditional independence formulation of causality equivalent to directed information [13], yet it did not refer to the necessary first step of causal deduction with observational data, which is deconfounding.

Directed information as a measure of strength of causal effect was criticised for vanishing in the presence of direct causal effect in the underlying Pearlian DAG [28,29] as well as for failing to detect the direction of the causal effect [30]. This critique correctly states that directed information alone is not a proper measure of the strength of causal effect, nevertheless it is rendered moot when one correctly interprets directed information as a measure of no confounding (we refer to it in Section 4).

Recently, directed information has been also applied to areas as diverse as learning hierarchical policies in reinforcement learning [31], modelling privacy loss in cloud-based control [32,33], learning polytrees for stock market data [34], submodular optimisation [35], EEG activity description [36], financial data exploration [37,38] and analysis of spiking neural networks [39]. New methods of directed information estimation [40] as well as generalisations to Polish spaces [41] have been proposed. All of this work, however, treats directed information as a measure of causality strength only and ignores its correct interpretation as measure of no confounding (i.e., the first step in the causal discovery procedure).

1.4. Related Work on Graphical Models for Causality

Causal relationships are frequently represented with graphical models, that is, sets of random variables depicted as nodes and relationships between them represented by different types of edges.

The basic goal of such models is to encode *dependence structures* of the underlying probability distribution with graph theoretical criteria such as d-separation [42]. When used in the context of causality modelling, one also makes sure that the connections between nodes have a causal interpretation, such as the *data generating process* of the underlying distribution for arrows [25].

A simple graphical model used for causality is the Pearlian DAG encoding both conditional independence relations and the causal data generating process with arrows. Capturing additional information about the dependence structure with the graph theoretic criterion was the motivation for more elaborate graphical models [43]. Completed Partially Directed Acyclic Graphs [9] allow both directed and undirected edges and describe equivalence classes of DAGs which encode the same conditional independence relations with d-separation. Ancestral graphs [44] extend the set of edges with bi-directed edges and allow one to model hidden and selection variables by assuring closure with respect to marginalisation and conditioning. Further extensions of ancestral graphs include maximal ancestral graphs and partial ancestral graphs [45], the latter being the output of popular structure learning algorithms such as FCI [9].

Another motivation for extending the simple DAG model stems from considering the data generating process rather than trying to encode more information about conditional independence relations alone. From the point of view of the data generating process, a DAG describes a set of relationships, where each variable is generated from the set of its parents and external noise [46,47]. *Chain graphs* are an extension of DAGs in which undirected edges between nodes are allowed as long as no semi-directed cycles (cycles with directed and undirected edges) emerge [48,49]. The corresponding data generating process consists of two levels. First, as in DAGs, every set of nodes connected with undirected edges (called *chain component*) depends on the set of all of its members' parents. Secondly, within every chain component, every node depends on all the other nodes without any specified direction of the dependence (which can be modelled as Gibbs sampling of the nodes in the chain component until the pdfs of the nodes reach an equilibrium). This interpretation of chain graphs' data generating process and Markov properties was proposed by Lauritzen and Wermuth [48,50,51] and later used as a basis for modelling causal relationships [52,53]. We build on this interpretation of chain graphs in Section 3.

Alternative ways of encoding conditional independence relations in chain graphs have been put forward [54] and [55]. Both have subsequently been extended to account for up to two edges between nodes and to exclude only directed cycles (instead of semi-directed ones) [56,57]. The resulting graphical models are called acyclic directed mixed graphs (ADMGs) [58]. Factorisation criteria along with corresponding algorithms have also been considered for both chain graphs and ADMGs [59,60].

1.5. Paper Contributions

In this paper, we make the following contributions:

- we formulate a two step procedure for causal deduction in two most widely known frameworks of causality modelling and show that the proposed information theoretic causal effect quantification is equivalent to it,
- we relate to various definitions of directed information and unify them within our approach,
- we clear some of the confusion persistent in previous attempts to information theoretic causality modelling.

The remainder of this paper is structured as follows. Section 2 introduces our method of causal deduction with information theoretic terms. Subsequently, we explain the existing differences between definitions of directed information, motivate chain graphs as a unifying structure and define directed information for chain graphs in Section 3. We relate to the critique of directed information in Section 4. We conclude with final remarks and an outline of future work in Section 5.

2. Proposed Method for Causal Effect Identification

In this section, we formalise the two-step causal deduction procedure with information theoretic terms outlined in Section 1.1. Recall that the two steps for quantifying the causal effect of one variable on another are:

S.1 Make sure that the variables are not confounded or find a set of variables confounding them.
S.2 Use the set found in S.1 (if it exists) to quantify the causal effect.

We elaborate on Step S.1 in the existing frameworks of causal deduction in Section 2.1.1 and in the proposed information theoretic framework in Section 2.2.1. Similarly, Step S.2 is described in Sections 2.1.2 and 2.2.2, respectively. First, we formally define the necessary concepts from the Pearlian and Neyman-Rubin potential outcome frameworks in Sections 2.1.1 and 2.1.2 and from information theory in Section 2.1.3.

2.1. Notation and Model Set-Up

Graphical models, in particular DAGs, are often employed for modelling causal relationships. Pearlian DAGs [25,61–63] represent both direct causal relationships between variables (expressed as arrows) and factorisation of the joint probability distribution of the variables (encoded as conditional independence relations).

A Pearlian DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{X_1, X_2, \ldots, X_n\}$ encodes conditional independence relations with d-separation (For the definition and examples of d-separation in DAGs, see textbooks by Lauritzen [42] or by Pearl [25], Chapters 1.2.3 and 11.1.2). This means that any pair of sets of variables in $\mathcal{V}$ d-separated by Z is conditionally independent given Z. The following probability factorisation and data generating process are assumed for a Pearlian DAG ([25], Chapter 3.2.1):

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i | pa(X_i)) \tag{1}$$

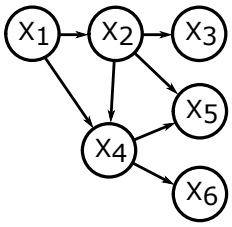
$$X_i = f_i(pa(X_i), U_i), \tag{2}$$

where $pa(X_i)$ stands for the set of direct parents of X_i and U_i are exogenous noise variables. If U_i are pairwise independent and each is independent of non-descendants of X_i, then the corresponding Pearlian DAG is called Markovian. If the joint distribution $P(U_1, U_2, \ldots, U_n)$ permits correlations between exogenous variables (which can be used, for example, to represent unmeasured common causes for elements of $\mathcal{V}$), the model is called semi-Markovian ([25], Chapter 3.2.1). The general information theoretic language for causality proposed in this paper remains valid for semi-Markovian and non-Markovian models, but we will confine ourselves to the Markov case in the current paper, since it suffices for describing the basic causal concepts such as the back-door criterion and the average causal effect.

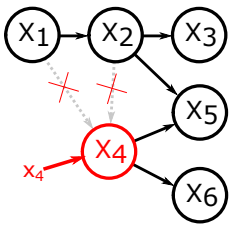
The causal meaning of a Pearlian DAG is formalised with the idea of an intervention: intervening on a variable or a set of variables means setting it to a preselected value and suppressing the influence of other variables on it. This results in the interventional distribution defined formally as ([25], Chapter 3.2.3):

$$P(X_1, X_2, \ldots, X_n | \mathrm{do}(X_i = x_i)) = \begin{cases} \prod_{j \neq i} P(X_j | pa(X_j)) & \text{if } X_i = x_i \\ 0 & \text{if } X_i \neq x_i. \end{cases} \tag{3}$$

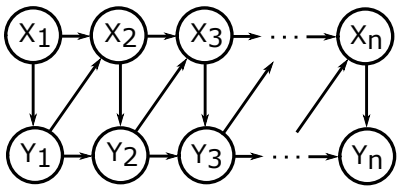
This definition formalises the motivating idea of an intervention by leaving out the term $P(X_i|pa(X_i))$ from the product. We will denote $\mathrm{do}(X_i) := \mathrm{do}(X_i = x_i)$ whenever it does not lead to confusion. Examples of Pearlian DAGs with interventions are given in Figure 1.



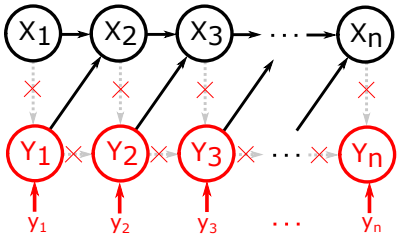
(a) $P(X_1, \ldots, X_6) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_1, X_2)P(X_5|X_2, X_4)P(X_6|X_4)$



(b) $P(X_1, \ldots, X_6|\mathrm{do}(X_4)) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ P(X_5|X_2, X_4)P(X_6|X_4)$



(c) $P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \prod_{i=1}^{n} P(X_i|X_{i-1}, Y_{i-1})P(Y_i|X_i, Y_{i-1})$



(d) $P(X_1, \ldots, X_n, Y_1, \ldots, Y_n|\mathrm{do}(Y_1, \ldots, Y_n)) = \prod_{i=1}^{n} P(X_i|X_{i-1}, Y_{i-1})$

Figure 1. Examples of interventions performed on directed acyclic graphs (DAGs) with resulting probability factorisations. Left: observational distributions and factorisations. Right: interventional distributions and factorisations.

The assumed functional characteristic of each child-parent relationship as defined in the data generating process of a Markovian Pearlian DAG (Equation (2)) encodes the same conditional independence relationships as the standard factorisation in Equation (1) [46]. Moreover, one can show that the *Causal Markov Condition* holds for a Markovian Pearlian DAG ([47], Theorem 1): the distribution defined by Equation (2) factorises according to Equation (1). Finally, the functional characteristic of all f_i along with its equivalence to the factorisation according to Equation (2) and the definition of intervention make it possible to formalise the concept of *modularity* [61] of Pearlian DAGs: for any node $X \in \mathcal{V}$, its conditional distribution given its parents does not depend on interventions on any other nodes in $\mathcal{V}$. When discussing Pearlian DAGs, we will also assume *positivity*: for any $X \in \mathcal{V}$ and a set $Z \subset \mathcal{V}$ of non-descendants of X, $P(X = x|Z) > 0$ with probability 1 for every x, that is, none of the modelled events have probability 0. Note that in the light of the above discussion, Pearlian graphs can be interpreted both as Bayesian networks imbued with a causal meaning and as structural equation models (Markovian models with non-parametric f_i in Equation (2)).

The counterpart of the intervention in the Neyman-Rubin causal model are the potential outcomes of a treatment. In the Neyman-Rubin causal model, potential outcomes $Y(0)$ and $Y(1)$ corresponding to a binary treatment variable X are equivalent to the interventional distributions of $P(Y|\mathrm{do}(X = 0))$ and $P(Y|\mathrm{do}(X = 1))$ [64–67]. Formally, for $X, Y \in \mathcal{V}$ and $X = \{0, 1\}$ being a binary variable, potential outcomes $Y(0)$ and $Y(1)$ are equal to the interventional distributions of $P(Y|\mathrm{do}(X = 0))$ and $P(Y|\mathrm{do}(X = 1))$ and variables X and Y in the potential outcomes model can be modelled as nodes in a Pearlian DAG [25].

Throughout the rest of this paper we will assume $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to be a Pearlian DAG as described above.

2.1.1. Controlling Confounding Bias

The interventional distribution can be computed directly whenever arbitrary interventions in the Pearlian DAG can be performed and measured. This corresponds to randomised treatment

assignment in the Neyman-Rubin potential outcome framework (e.g., assigning patients randomly to treatment and control groups such that the assignment does not depend on any other variables in the model). The goal of the first step in the procedure of quantification of causal effects (S.1 in Section 2) is to establish if and how it is possible to circumvent the necessity of measuring the interventional distribution or performing a randomised experiment. This is done by searching for a set of variables which make it possible to express the interventional distribution with observational distributions only.

Given the Pearlian DAG $\mathcal{G}$ with observational data only (i.e., a sample from a subset of the nodes of the Pearlian DAG), one can specify conditions under which interventional distribution $P(Y|\mathrm{do}(X))$ can be derived [25,68]. If all parents of X are measured, it can be shown that Equation (3) can be transformed to the following form (*adjusting for direct causes*, that is, parents of X in $\mathcal{G}$) ([25], Theorem 3.2.2):

$$P(Y|\mathrm{do}(X)) = \sum_{X' \in pa(X)} P(Y|X, X')P(X'). \tag{4}$$

The procedure of Equation (4), that is, conditioning on a set of variables and then averaging by the probability of this set is referred to as *adjusting* and the said set is called the *adjustment set*. This leads to the following general definition.

Definition 1 (Adjustment set [69])**.** *In a Pearlian DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, for pairwise disjoint $X, Y, Z \subseteq \mathcal{V}$, Z is an adjustment set relative to the ordered pair (X, Y) if and only if:*

$$P(Y|do(X)) = \sum_{Z' \in Z} P(Y|X, Z')P(Z') = \mathbb{E}_Z \left[Y|X, Z\right]. \tag{5}$$

The point of adjusting is to remove spurious correlations between X and Y while not introducing new ones. In this light, controlling confounding bias amounts to finding a set Z such that, upon adjusting for Z, $P(Y|\mathrm{do}(X))$ can be computed from observational data. Equation (4) shows that the set of parents of X is such a set. This result has been generalised thus making it possible to find adjustment sets also in the case where not all variables in the Pearlian DAG are measured. Such adjustment sets must fulfil the back-door criterion. The generalisation is thus called the back-door adjustment.

Definition 2 (Back-door criterion, [25], Definition 3.3.1)**.** *A set of variables $Z \subseteq \mathcal{V}$ satisfies the back-door criterion relative to an ordered pair of variables (X, Y) if it fulfils the two following conditions:*

- *no node in Z is a descendant of X and,*
- *Z blocks (d-separates) all paths between X and Y that contain an arrow into X.*

$Z \subseteq \mathcal{V}$ satisfies the back-door criterion relative to a pair of disjoint subsets of $\mathcal{V}$, $(\mathcal{X}, \mathcal{Y})$ if it satisfies the back-door criterion relative to any pair (X, Y) with $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.

Note that the first condition in Definition 2 is equivalent to Z being a set of post-treatment variables or covariates, that is, variables not affected by treatment in the Neyman-Rubin potential outcomes model [64,70].

Theorem 1 (Back-door adjustment, [25], Theorem 3.3.2)**.** *Let $X, Y, Z \subseteq \mathcal{V}$ be disjoint. If Z satisfies the back-door criterion relative to the pair (X, Y), then it is an adjustment set relative to this pair.*

Examples of adjustments sets corresponding to adjusting for direct causes and the back-door criterion are presented in Figure 2.

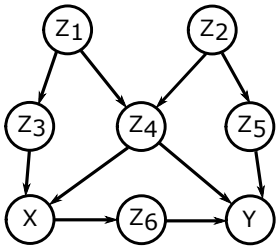


Figure 2. Adjustment sets for (X, Y). Back-door adjustment [25]: $\{Z_3, Z_4\}$ and $\{Z_4, Z_5\}$ satisfy the back-door criterion with respect to (X, Y). Only the former corresponds to adjusting for direct causes of X.

The observation that adjusting for a set of variables means the removal of spurious correlations without introducing new ones leads to the following definition of no confounding [25,68]:

Definition 3 (No confounding, [25], Definition 6.2.1)**.** *In a Pearlian DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, an ordered pair (X, Y) with $X, Y \subset \mathcal{V}$, $X \cap Y = \emptyset$ is not confounded if and only if $P(Y = y|do(X = x)) = P(Y = y|X = x)$ for all x, y in their respective domains.*

In the context of the Neyman-Rubin potential outcome model, one often deals with confounding by assuming strong ignorability given Z [70]: $\{Y(0), Y(1)\} \perp\!\!\!\perp X|Z$. It can be shown that strong ignorability implies that Z satisfies the back-door criterion by constructing an appropriate Pearlian DAG ([25], Chapter 11.3.2).

2.1.2. Quantifying Causal Effects

In an *unconfounded* setting, that is, after all spurious correlations between cause and effect have been adjusted for (in the cases where it is possible), one can proceed to quantify the strength of the remaining causal effect (i.e., Step S.2 in Section 2). In a Pearlian DAG, the interventional distribution $P(Y|\text{do}(X))$ describes the causal effect of X on Y. Note that, as described in Section 2.1.1, this distribution can be computed from observational data when an appropriate adjustment set of variables has been measured, be it all direct ancestors of X or variables satisfying the back-door criterion. In the Neyman-Rubin potential outcome framework, one of the most common measures of causal strength for binary treatments is the average causal effect, also referred to as average treatment effect [67,71]:

Definition 4 (Average Causal Effect [71])**.** *Let X be a binary treatment variable and $Y(1)$ and $Y(0)$ stand for potential outcomes corresponding to the counterfactuals $X = 1$ and $X = 0$, respectively. Define:*

$$ACE(X, Y) = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|do(X = 1) - Y|do(X = 0)]. \tag{6}$$

An equivalent of ACE restricted to a subspace of the population with a given value of a certain variable Z which is a non-descendent of the treatment variable X can be defined [68,72,73].

Definition 5 (Specific Causal Effect, [72], Definition 9.1)**.** *Let X be a binary treatment variable and $Y(1)$ and $Y(0)$ stand for potential outcomes corresponding to the counterfactuals $X = 1$ and $X = 0$, respectively. Let $Z \subseteq \mathcal{V}$ be a set of non-descendants of X. Define:*

$$SCE(X, Y) = \mathbb{E}[Y|do(X = 1), Z = z - Y|do(X = 0), Z = z]. \tag{7}$$

The Specific Causal Effect can be thought of as ACE conditional on a particular value of $Z = z$ (also defined as Conditional Average Causal Effect [74,75]) with the additional requirement that Z is a non-descendant (or a set of non-descendants) of X in the underlying Pearlian DAG.

Clearly, ACE is a function of two interventional distributions $P(Y|\text{do}(X = 1))$ and $P(Y|\text{do}(X = 0))$. In general, ACE requires the observation of both potential outcomes. In can be shown however that, when a set of variables Z satisfies the back-door criterion with respect to X and $\{Y(0), Y(1)\}$ or strong ignorability is assumed, ACE is estimable from observational data [25,68,73]:

$$\begin{aligned} ACE(X,Y) &= \mathbb{E}_{P(Z)}\big[\mathbb{E}[Y(1)|Z] - \mathbb{E}[Y(0)|Z]\big] \\ &= \mathbb{E}_{P(Z)}\big[\mathbb{E}[Y|\text{do}(X=1),Z] - \mathbb{E}[Y|\text{do}(X=0),Z]\big] \\ &= \mathbb{E}_{P(Z)}\big[\mathbb{E}[Y|X=1,Z] - \mathbb{E}[Y|X=0,Z]\big]. \end{aligned} \tag{8}$$

This corresponds to averaging over the SCE given all the possible values of Z or over the conditional average causal effect and mirrors the adjustment formula from Definition 1 and Theorem 1.

Despite its simplicity and limitation to the binary case ACE remains one of the most popular measures of causal effect because of its interpretability (for example in the medical setting where it quantifies the effect of a particular treatment strategy).

2.1.3. Information Theory and Directed Information

We now provide definitions of the necessary concepts from information theory as well as the general definition of directed information. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a Pearlian DAG and assume $X, Y \subseteq V$.

Define the *Kullback-Leibler divergence* between two (discrete or continuous) probability distributions P and Q as $D_{KL}(P(X) \,||\, Q(X)) = \mathbb{E}_{P(X)} \log \frac{P(X)}{Q(X)}$ and the *conditional Kullback-Leibler divergence* as $D_{KL}(P(Y|X) \,||\, Q(Y|X) \mid P(X)) = \mathbb{E}_{P(X,Y)} \log \frac{P(Y|X)}{Q(Y|X)}$

The *mutual information* between X and Y is then defined as $I(X;Y) = D_{KL}(P(X,Y) \,||\, P(X)P(Y))$ and the *conditional mutual information* given Z as $I(X;Y|Z) = D_{KL}(P(X,Y,Z) \,||\, P(X|Z)P(Y|Z)P(Z))$.

Let $H\left[P(X)\right] = -\mathbb{E}_{P(X)}\left[\log P(X)\right]$ denote *entropy* for discrete and *differential entropy* for continuous X. Analogously, $H\left[P(X|Y)\right] = -\mathbb{E}_{P(X,Y)}\left[\log P(X|Y)\right]$ denotes *conditional entropy* for discrete and *conditional differential entropy* for continuous X and Y. For discrete variables, define additionally $H\left[P(X|Y=y)\right] = -\mathbb{E}_{P(X|Y=y)}\left[\log P(X|Y=y)\right]$ so that the following holds: $H\left[P(X|Y)\right] = \mathbb{E}_{P(Y)}\left[H\left[P(X|Y=y)\right]\right] = -\mathbb{E}_{P(X,Y)}\left[\log P(X|Y)\right]$.

As pointed out in Section 1.3, several definitions of directed information have been proposed in the literature. We adopt the definition of directed information given in [15]. In Section 3 we show that this definition subsumes other definitions for time series.

Definition 6 (Directed Information [15]). *Let $X, Y \subseteq \mathcal{V}$ be disjoint.*

$$I(X \to Y) = D_{KL}\big(P(X|Y) \,||\, P(X|do(Y)) \mid P(Y)\big) = \mathbb{E}_{P(X,Y)} \log \frac{P(X|Y)}{P(X|do(Y))} \tag{9}$$

One might also consider interventional distribution with conditioning on a set of passive observations. This leads to the definition of conditional directed information [15] for three disjoint sets $X, Y, Z \subseteq V$.

Definition 7 (Conditional Directed Information [15]). *Let $X, Y, Z \subseteq \mathcal{V}$ be pairwise disjoint.*

$$I(X \to Y|Z) = D_{KL}\Big(P(X|Y,Z) \,||\, P(X|do(Y),Z) \mid P(Y,Z)\Big) = \mathbb{E}_{P(X,Y,Z)} \log \frac{P(X|Y,Z)}{P(X|do(Y),Z)} \tag{10}$$

Note that the expression $P(X|do(Y), Z)$ means conditioning on Z in the interventional distribution $P(X|do(Y))$ as defined in Equation (3) (i.e., the intervention $do(Y)$ is performed before conditioning on Z). In particular,

$$P(X|do(Y), Z) = \frac{P(X, Z|do(Y))}{P(Z|do(Y))}. \tag{11}$$

Thus, conditional directed information compares the effect of conditioning on Z in two distributions: observational $P(X|Y)$ and interventional $P(X|\mathrm{do}(Y))$.

2.2. Causal Deduction with Information Theory

We now proceed to lay out the two-step procedure for *information theoretic* causal effect quantification. It consists of ensuring that the two sets of random variables between which the causal effect is to be identified are not confounded (possibly given an adjustment set) and subsequently quantifying the causal effect. The former step Step S.1 in Section 2 is achieved with (conditional) directed information; the latter (Step S.2 in Section 2) with (conditional) mutual information.

2.2.1. Controlling Confounding Bias with (Conditional) Directed Information

In the first step of the information theoretic causal effect quantification procedure one checks whether the two variables of interest are not confounded and if they are, whether any set Z can serve as adjustment set. It is straightforward to note that the definition of directed information $I(X \to Y)$ provided in Definition 6 is equivalent to the criterion for no confounding between (Y, X) (Definition 3). This is formalised in Proposition 1.

Proposition 1. *An ordered pair (X, Y) with $X, Y \subseteq \mathcal{V}$, $X \cap Y = \emptyset$ is not confounded if and only if $I(Y \to X) = 0$.*

The extension of this basic result to the case of adjusting for confounding bias with the back-door criterion was formulated in [15]:

Proposition 2 (Theorem 1 in [15])**.** *Let $Z \subset \mathcal{V}$ be a set of non-descendants of X and let $X \cap Y = \emptyset$. Then: Z is an adjustment set for the pair (X, Y) if and only if $I(Y \to X|Z) = 0$.*

Propositions 1 and 2 formalise the interpretation of directed information: if the (conditional) directed information from Y to X vanishes, the causal effect of X on Y is identifiable with observational data, possibly after adjusting for the conditioning set Z. If directed information is greater than 0, performing an intervention on X has influenced the distribution of Y, hence the difference must stem from the connections between X and Y in $\mathcal{V}$, which were destroyed while intervening on X (such connections correspond to Z satisfying the back-door criterion). Note that for the identification of the causal effect $X \to Y$, the 'inverse' directed information $I(Y \to X)$ must vanish.

The interpretation of directed information as a measure of no confounding explains the misunderstandings in situations where directed information (or Granger causality, transfer entropy) is used to quantify direct causal influence. We relate to such 'counterexamples' in Section 4.

2.2.2. Quantifying the Causal Effect with (Conditional) Mutual Information

As shown in Section 2.2.1, $I(Y \to X) = 0$ implies that the causal effect of X on Y can be identified with observational data, for example, according to (Theorem 1 and Equation (8)). We now show that in this unconfounded setting, (conditional) mutual information captures the causal effect in a manner analogous to the average causal effect.

Quantifying the causal effect of an intervention with an interpretable value requires proposing a meaningful functional summarising the difference between two (or more) distributions. In the Pearl framework, the causal effect is defined as a function from X to $P(Y|\text{do}(X))$, so it captures full distributional information about all possible interventions setting X to different values. It therefore represents all available information but is difficult to interpret since it consists of a continuous space of probability distributions. In the Neyman-Rubin causal model, the ACE (Definition 4) makes use of the fact that X is binary and reduces both resulting distributions to their means.

We prove that by taking the middle ground, one can meaningfully quantify the causal effect with mutual information and conditional mutual information in an unconfounded setting. To this end, we employ the weighted Jensen-Shannon divergence [76,77], which is sensitive to more than just the first moment of a distribution, as a measure of difference between interventional distributions. We then show that SCE and ACE (Definitions 4 and 5) are equivalent to conditional mutual information and mutual information, respectively, when the difference of means is replaced with the Jensen-Shannon divergence.

Definition 8 (Weighted Jensen-Shannon Divergence (JSD) [76])**.** *Let p, q be probability distributions and $\pi_q, \pi_r \in \mathbb{R}_+ \cup \{0\}$ be weights with $\pi_q + \pi_r = 1$. The weighted Jensen-Shannon divergence (JSD) is defined as:*

$$JSD(q \,||\, r) = H[\pi_q q + \pi_r r] - \pi_q H[q] - \pi_r H[r]. \tag{12}$$

Note that JSD is sometimes equivalently defined for $\pi_q = \pi_r = \frac{1}{2}$ as symmetrised Kullback-Leibler divergence between p, q and $m := \frac{1}{2}(p+q)$: $JSD(p,q) = \frac{1}{2}(D_{KL}(p \,||\, m) + D_{KL}(q \,||\, m))$ [77]. JSD has recently been applied in many machine learning areas such as GANs [78], bootstrapping [79], time series analysis [80] or computer vision [81].

We first show that for two sets of variables which are not confounded, mutual information quantifies the Jensen-Shannon divergence between two interventional distributions corresponding to the application of a treatment and lack thereof (see Appendix A for the proof).

Proposition 3 (Quantifying causal effects with mutual information)**.** *Assume an ordered pair (X, Y) in a Pearlian DAG with $X, Y \subseteq \mathcal{V}$, $X \cap Y = \emptyset$ and denote the interventional distributions and corresponding weights as follows:*

$$\begin{aligned} q &= P(Y|do(X=1)), \quad \pi_q = P(X=1) \\ r &= P(Y|do(X=0)), \quad \pi_r = P(X=0). \end{aligned} \tag{13}$$

Then the following holds:

$$\text{if } I(Y \to X) = 0, \text{ then } I(X;Y) = JSD(r \,||\, q).$$

We now proceed to show that when two sets of variables are confounded, but a third set satisfying the back-door criterion relative to these two sets exists, Jensen-Shannon divergences between interventional distributions conditioned on a particular value of the third set and averaged over all values of this set are equal to a KL divergence and conditional mutual information, respectively. These divergences are analogous to SCE and ACE with differences of means replaced with JSD.

Proposition 4 (Quantifying specific causal effects)**.** *Assume an ordered pair (X, Y) in a Pearlian DAG with $X, Y \subseteq \mathcal{V}$, $X \cap Y = \emptyset$ and $Z \subset \mathcal{V}$ which satisfies the back-door criterion (Definition 2).*
Denote the interventional distributions and corresponding weights for a given value of $Z = z$ as follows:

$$\begin{aligned} q_z &= P(Y|do(X=1), Z=z), \quad \pi_{q_z} = P(X=1|Z=z) \\ r_z &= P(Y|do(X=0), Z=z), \quad \pi_{r_z} = P(X=0|Z=z). \end{aligned} \tag{14}$$

Then the following holds:

$$\text{if } I(Y \to X|Z) = 0, \text{ then } JSD(r_z \,||\, q_z) = D_{KL}\Big(P(X,Y|Z=z)||P(X|Z=z)P(Y|Z=z)\Big).$$

The proof is provided in Appendix A. In fact, it suffices that the equivalent of conditional directed information for the particular z vanishes: $I(X \to Y|Z=z) := D_{KL}(P_{X|Y,Z=z}||P_{X|\mathrm{do}(Y),Z=z}|P_{Y,Z=z}) = \mathbb{E}_{P_{X,Y|Z=z}} \log \frac{P(X|Y,Z=z)}{P(X|\mathrm{do}(Y),Z=z)}$.

The following Corollary justifies using conditional mutual information as a measure of causal effect in an unconfounded setting (see Appendix A for the proof).

Corollary 1. *(Quantifying average causal effects with conditional mutual information) Assume an ordered pair (X,Y) in a Pearlian DAG with $X,Y \subseteq \mathcal{V}$, $X \cap Y = \emptyset$ and $Z \subset \mathcal{V}$ which satisfies the back-door criterion (Definition 2). Denote the interventional distributions and corresponding weights q_z, r_z, π_{q_z}, π_{r_z} as in Equation (14) in Proposition 4.*

Then the following holds:

$$\text{if } I(Y \to X|Z) = 0, \text{ then } \mathbb{E}_Z[JSD(r_z \,||\, q_z)] = I(X;Y|Z).$$

Propositions 3 and 4 and Corollary 1 justify using mutual information and conditional mutual information for quantifying causal effects of X on Y in unconfounded settings (i.e., whenever X and Y are not confounded or a set Z satisfying the back-door criterion exists). This corresponds to Step S.2.

Both directed information and conditional mutual information have been proposed as measures of quantifying causal effects. Both measures have also been criticised for their shortcomings in the ability of capturing these effects [28–30,82,83]. In this section we showed that only their combination yields a rigorous framework for causal effect quantification in Pearlian DAGs. Table 1 summarises our approach.

Table 1. Summary of information theoretic causal effect quantification and comparison of the two steps to the Pearl and Neyman-Rubin potential outcome frameworks.

	Pearlian Framework	**Neyman-Rubin Potential Outcome Framework**	**Information Theoretic Framework**
Ensuring no confounding	back-door criterion	strong ignorability	conditional directed information $= 0$
Causal effect quantification	interventional distribution	average causal effect	conditional mutual information

3. Unification of Existing Approaches for Time Series

As stated in Section 1.3, before its general formulation given in Definitions 6 and 7, directed information was defined for discrete channels (or, equivalently, time series) [11,16]. This has resulted in the situation where two competing definitions of directed information for time series are in use: with and without incorporating the instantaneous point in the other time series, that is, with or without 'conditioning on the present'. Denote a set of n ordered variables in a Pearlian (a time series with n time points) DAG as $X^n := X_1, X_2, \ldots, X_n$. Formally, directed information between time series X^n and Y^n was defined as:

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}) \tag{15}$$

by Massey [11] and adopted in this form by some authors [19–21] with the justification that X^n and Y^n are "synchronised" and X_i and Y_i "occur at the same time" ([19], Chapter 3.1.1). In parallel, the following definition of directed information was put forward in [13,16,22,23] with the argument

that "since the causation is already known [...], it is notationally convenient to use synchronised time" [13]:

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^{i-i}; Y_i | Y^{i-1}). \tag{16}$$

Moreover, definitions on different levels of generality are present varying from two and multiple time series as above to general DAGs as in Definitions 6 and 7. In this section, we show that both discrepancies vanish when one considers the different definitions of directed information as special cases of Definitions 6 and 7. We thus unify various formulations of directed information and conditional directed information into one.

To this end, we first show in Section 3.1 that the two variants of directed information for time series defined in Equations (15) and (16) are indeed special cases of Definition 6 for different Pearlian DAGs corresponding to different intuitive assumptions concerning time ordering. We subsequently extend the DAGs with a third, confounding, time series and derive formulas for conditional directed informations for these DAGs according to Definition 7.

We then relate the reason for the discrepancy between conditioning on the present and lack thereof to the motivation of using chain graphs in causality modelling and introduce chain graphs in Section 3.2.

Note that directed information for a general Pearlian DAG with a given ordering can be obtained by comparing factorisations of the observational and interventional DAGs [15]. Indeed, expressing Definition 6 as

$$I(X \to Y) = \mathbb{E}_{P(X,Y)} \log \frac{P(X|Y)}{P(X|\mathrm{do}(Y))} = \mathbb{E}_{P(X,Y)} \log \left[\frac{P(X,Y)}{P(X|\mathrm{do}(Y))P(Y)} \right] \tag{17}$$

results in the observational distribution in the numerator and a product of the interventional distribution and the marginal distribution of the variables intervened upon in the denominator. Factorisations of both distributions can be directly read off the corresponding DAGs. Different forms for directed mutual informations result from the different orderings imposed on the underlying Pearlian DAGs.

3.1. Directed Information for Time Series Represented with DAGs.

We now show that the definitions of directed information for time series Equations (15) and (16) are special cases of Definition 6. We do this by defining appropriate Pearlian DAGs (corresponding to *full time ordering* and *partial time ordering*) and applying Definitions 6 and 7 as well as factorisations of observational and interventional distributions (Equations (1) and (3)) to them.

Consider a Pearlian DAG $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E})$, where $|\mathcal{V}| = 2n$ and a total order on $\mathcal{V}$ is given. This means that $\mathcal{V} = (V_1, V_2, \ldots, V_{2n-1}, V_{2n})$, with $\mathcal{E}$ consisting of all possible arrows pointing to the future, that is, $V_i \to V_j$ with $i < j$. Now, define $X^n = (X_1, X_2, \ldots, X_n) = (V_1, V_3, \ldots, V_{2n-1})$ and $Y^n = (Y_1, Y_2, \ldots, Y_n) = (V_2, V_4, \ldots, V_{2n})$. DAG $\mathcal{G}_1$ is depicted in Figure 3. Theorem 2 shows the formula for directed information that follows from applying Definition 6 to $\mathcal{G}_1$.

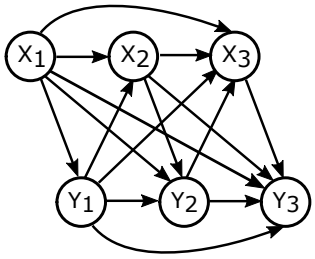


Figure 3. Pearlian DAG $\mathcal{G}_1$ representing *full ordering* considered in Theorem 2.

Theorem 2. *In the Pearlian DAG $\mathcal{G}_1$ directed information from X^n to Y^n has the following form:*

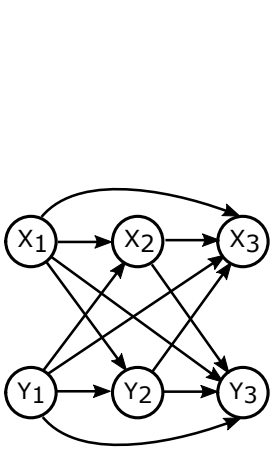
$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}). \tag{18}$$

In the same DAG $\mathcal{G}_1$, directed information from Y^n to X^n has the following form:

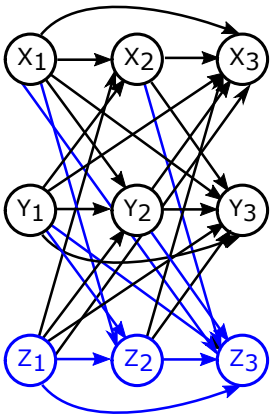
$$I(Y^n \to X^n) = \sum_{i=1}^{n} I(Y^{i-1}; X_i | X^{i-1}). \tag{19}$$

See Appendix A for the proof. Note that Equation (18) is indeed equivalent to the directed information defined on time series in [11] (Equation (15)).

Now consider a Pearlian DAG $\mathcal{G}_2$ similar to $\mathcal{G}_1$ ($\mathcal{G}_2 = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$) but with a slight twist. Let now X^n and Y^n be aligned, that is, indexed at the same time points. Let $\mathcal{E}$ again consist of all possible arrows pointing to the future (i.e., all arrows $X_i \to X_j$, $Y_i \to Y_j$, $X_i \to Y_j$, $Y_i \to X_j$, with $i < j$). $\mathcal{G}_2$ is shown in Figure 4a. Applying Definition 6 to $\mathcal{G}_2$ as well as $\mathcal{G}_2$ together with a third, confounding, time series (Figure 4b) yields Theorem 3.



(**a**) Pearlian DAG $\mathcal{G}_2$.



(**b**) The same time series along with a third, *confounding* time series Z (in blue).

Figure 4. Pearlian DAGs representing *partial ordering* considered in Theorem 3.

Theorem 3. *In the Pearlian DAG $\mathcal{G}_2$ directed information from X^n to Y^n has the following form:*

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}). \tag{20}$$

Conditioning on an aligned time series Z^n (see Figure 4b) yields:

$$I(X^n \to Y^n | Z^n) = \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}, Z^{i-1}). \tag{21}$$

See Appendix A for the proof. Analogously to Theorem 2, Equation (20) is equivalent to the directed information defined on time series in [16] (Equation (16)).

3.2. Factorisations and Interventions in Chain Graphs.

In Section 3.1 we showed that two definitions of directed information proposed in the literature are subsumed by Definition 6. These two definitions differ in how they treat events that are supposed to be time-aligned. It is therefore not clear what causal assumptions or hypotheses should be allowed to model such events: if an association is observed between them, can it be explained by a directed

arrow in the data generating process in Equation (2) (and if so, which direction should be assumed), by an unmeasured variable in a semi-Markovian model or can it only be an artefact of the functional form of the other arrows?

Similar considerations have led to the extension of DAGs to chain graphs as graphical models for causality. Potential presence of associations between variables which cannot be attributed to an underlying causal process (e.g., because the direction of causality cannot be established with available measurements, there exists an unmeasured confounding variable or a feed-back mechanism) motivated a causal interpretation of chain graphs [52,53] analogous to the causal interpretation of DAGs introduced in Section 2.1. The said non-causal direct associations are modelled with undirected adges between variables.

A chain graph (CG) $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is an extension of DAG in which $\mathcal{E}$ can also contain undirected edges and where no semi-directed cycles (i.e., cycles with directed and undirected edges) are allowed. This induces a new relationship between the elements of $\mathcal{V}$, distinct from parenthood: $X, Y \in \mathcal{V}$ are called *neighbours* if they are connected by an undirected edge. The set $\mathcal{T}$ of connected components (neighbours) of $\mathcal{V}$ obtained by removing all directed edges in a chain graph is called the set of chain components. In particular, chain graphs with no undirected edges or where all chain components are singletons are DAGs.

Analogously to Equations (2) and (3), the data generating process as well as interventional distribution have been defined for CGs. We follow the approach put forward in [52,53].

The data generating process of a CG is, again, an extension of that of a DAG (Equation (2)). As mentioned in Section 1.4, it consists of two levels. First, functional relationships of each child-parent pair of chain components are modelled: $\tau = f_\tau(pa(\tau), U_\tau)$ where $pa(\tau) = \bigcup_{X \in \tau} pa(X) \setminus \tau$. This corresponds to a DAG of all the chain components $\tau \in \mathcal{T}$. Secondly, for every chain component τ, a sampling procedure represented by g_τ is performed ([52], Section 6.3):

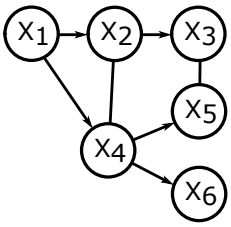
$$\tau = g_\tau(pa(\tau)). \tag{22}$$

here, g_τ represents the sampling function of the undirected graph τ. It takes all parents of τ as input and for every $X \in \tau$, it samples from its current distribution given $pa(\tau) \cup \tau \setminus \{X\}$ until reaching an equilibrium.

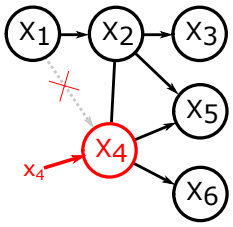
Just like the data generating process for DAGs motivates the definition of interventional distribution for DAGs (Section 2.1 and Equation (3)), the same reasoning can be applied to CGs, which leads to the following definition of the interventional distribution in a CG ([52], Section 6.4):

$$P(X|\text{do}(Y)) = \prod_{\tau \in \mathcal{T}} P(\tau \setminus Y | pa(\tau), \tau \cap Y). \tag{23}$$

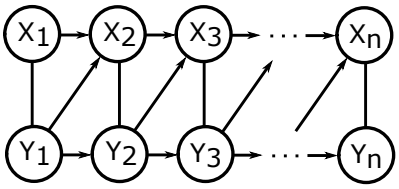
Thus, for every chain component τ that intersects with Y, $\tau \cap Y$ is removed from the factorisation (just like $P(X_j|pa(X_j))$ is removed from DAGs in Equation (3)) but still influences the remainder of the chain component τ by conditioning it. Examples of interventions in chain graphs are presented in Figure 5.



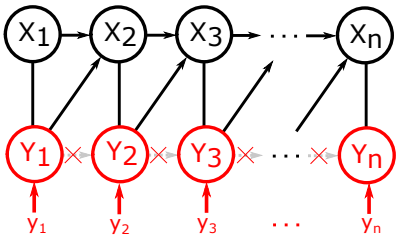
(**a**) $P(X_1, \ldots, X_6) =$
$P(X_1)P(X_2, X_4|X_1)P(X_3|X_2)\ P(X_5|X_2, X_4)P(X_6|X_4)$



(**b**) $P(X_1, \ldots, X_6|\mathrm{do}(X_4)) =$
$P(X_1)P(X_2|X_1, X_4)P(X_3|X_2)\ P(X_5|X_2, X_4)P(X_6|X_4)$



(**c**) $P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) =$
$\prod_{i=1}^{n} P(X_i, Y_i|X_{i-1}, Y_{i-1})$



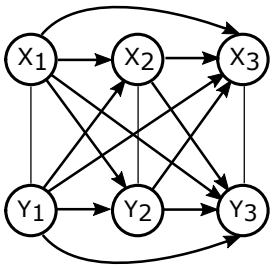
(**d**) $P(X_1, \ldots, X_n, Y_1, \ldots, Y_n|\mathrm{do}(Y_1, \ldots, Y_n)) =$
$\prod_{i=1}^{n} P(X_i|X_{i-1}, Y_{i-1}, Y_i,)$

Figure 5. Examples of interventions performed on chain graphs with resulting probability factorisations. Left: observational distributions and factorisations. Right: interventional distributions and factorisations. Note that, as opposed to Figure 1, $\{X_2, X_4\}$ (Figure 5a,b) and $\{X_i, Y_i\}$ (Figure 5c,d) form chain components.

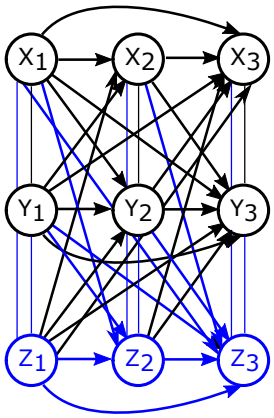
3.3. Directed Information for Chain Graphs Representing Aligned Time Series.

We now revisit directed information for time series motivated in Section 3.1. We first showed that two versions of directed information present in the literature (Equation (15) and (16)) are subsumed by Definition 6 and that the difference in motivations for the two versions is captured by the causal interpretation of chain graphs (Section 3.2). In this section, we propose to model aligned time series explicitly with chain graphs.

To this end, we define chain graph $\mathcal{H}_1 = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are as in DAG $\mathcal{G}_2$ from Theorem 3 and Figure 4a with $\mathcal{E}$ extended by additional undirected edges between every pair of X_i and Y_i. Thus, all sets $\{X_i, Y_i\}$ are chain components. $\mathcal{H}_1$ is depicted in Figure 6a.



(**a**) Chain graph $\mathcal{H}_1$



(**b**) The same time series along with a third, *confounding* time series Z (in blue).

Figure 6. Chain graphs considered in Theorem 4.

Theorem 4 shows the formula for directed information as well as conditional directed information in chain graphs presented in Figure 6.

Theorem 4. *In the chain graph $\mathcal{H}_1$ directed information from X^n to Y^n has the following form:*

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}). \tag{24}$$

Conditioning on an aligned time series Z^n (see Figure 6b) yields:

$$I(X^n \to Y^n | Z^n) = \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}, Z^{i-1}). \tag{25}$$

The proof, again, uses Definitions 6 and 7 and appropriate factorisations of observational and interventional distributions for chain graphs as defined in Equations (22) and (23) (see Appendix A).

4. Relation to Critique of Previous Information Theoretic Approaches

Directed information has been subject to criticism in the literature [28–30,84]. It concerned the time series formulation (as in Equations (15) and (16)), also in the form of transfer entropy or information flow. In the latter two forms, only the last term of the sum in Equations (15) and (16) is taken as the definition of directed information. All of the critique amounted to constructing examples where directed information fails to mirror intuitions or postulates concerning causal effect quantification. These postulates, however, are usually based on the erroneous assumption that directed information is by definition a measure of causal influence. As we described in Section 2, directed information is a measure of no confounding and constitutes the first step in the two-step procedure of causal effect quantification. In this section, we refer to the most common point of criticism raised in recent literature and show that it becomes irrelevant when one interprets directed information correctly and proceeds according to the information theoretic causal quantification procedure we proposed in Section 2.

Ay and Polani [28] consider a Pearlian DAG depicted in Figure 7. They note that transfer entropy from X to Y (i.e., directed information $I(X^{n-1} \to Y_n)$, defined by them as $I(X^{n-1}; Y_n | Y^{n-1})$) vanishes even though, intuitively, X directly influences Y (the example is symmetric in X and Y). Specifically, if one defines all the arrows in Figure 7 as noisy copy operations (i.e., one assumes $X_i = Y_{i-1} + \epsilon_{X_i}$ and $Y_i = X_{i-1} + \epsilon_{Y_i}$ with all $\epsilon \sim \mathcal{N}(0, \sigma^2)$ as Equation (2) in the underlying Pearlian DAG), then $I(X^{n-1} \to Y_n)$ decreases to 0 as $\epsilon \to 0$. The same critique was repeated by other authors [29,82]. It can, however, be easily explained with the two step method proposed in Section 2.

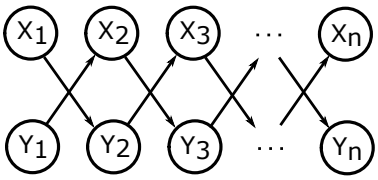


Figure 7. Example of "vanishing directed information" [28,29,82].

According to step S.1 of the causal effect quantification procedure described in Section 2, if one is interested in the causal effect of X^{n-1} on Y_n, one needs to first analyse the directed information $I(Y_n \to X^{n-1})$, since it measures whether the pair (X^{n-1}, Y_n) is not confounded. If $I(Y_n \to X^{n-1}) = 0$ in the underlying DAG, one can proceed to quantifying the causal effect with mutual information $I(Y_n; X^{n-1})$ (Step S.2).

Having established that, note that $I(Y_n \to X^{n-1})$ in the DAG from Figure 7 is indeed equal to 0:

$$\begin{aligned} I(Y_n \to X^{n-1}) &= \mathbb{E}_{P(X,Y)} \log \frac{P(X^{n-1}, Y_n)}{P(Y_n|\mathrm{do}(X^{n-1}))P(X^{n-1})} \\ &= \mathbb{E}_{P(X,Y)} \log \frac{P(Y_n|X_{n-1}) \prod_{i=2}^{n-1} P(X_i|X_{i-2})\ P(X_1)}{P(Y_n|X_{n-1})P(X_1) \prod_{i=2}^{n-1} P(X_i|X_{i-2})} = 0. \end{aligned} \tag{26}$$

Therefore, in order to clarify the criticism of directed information formulated in [28–30], it is essential to:

- use the directed information $I(Y_n \to X^{n-1})$ as a measure of no confounding,
- calculate $I(Y_n \to X^{n-1})$ according to the underlying DAG presented in Figure 7 (Equation (26)).

5. Conclusions

In this paper, we have proposed an attempt to bridge the most popular frameworks of causality modelling with information theory. To this end, we described a two step procedure of causal deduction, consisting of identifying confounding variables and subsequently quantifying the causal effect in an unconfounded setting, in each of these frameworks. We then expressed this procedure with infromation theoretic tools. Subsequently, we unified different definitions of directed information and clarified some of the confusion surrounding its causal interpretation. This is relevant since previous approaches to interpreting directed information were largely limited to the setting of time series and erroneously attributed causal effect quantification to directed information.

The full information theoretic description of causal deduction can be of interest to two communities. Firstly, for the statistical and causality community, since it provides a direct translation to the language of information theory, which has made inroads into machine learning recently. Secondly, it allows for the use of information theoretic machine learning models, such as the variational auto-encoder [85,86], deep information bottleneck [87,88], InfoGAN [89], and so forth, for causality modelling and integrating causal deduction in such models. The latter approach has already sparked interest in recent machine learning literature, for example, in the context of using causal relationships to facilitate transfer learning in deep models[8,90], explaining deep generative models and making them more interpretable [91,92] and boosting the performance of deep neural networks [93].

Future work includes elucidating information theoretic equivalents of further causal concepts such as the effect of treatment on the treated, propensity score based methods or double robustness models.

Author Contributions: Conceptualisation, A.W.; supervision, V.R.

Funding: This research was partially funded by the Swiss National Science Foundation grant number CR32I2159682.

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

Appendix A. Proofs

We first give proofs of Propositions 3 and 4 and Corollary 1.

Proof of Proposition 3.

$$\begin{aligned} JSD(r \,||\, q) &= JSD\Big(P(Y|\mathrm{do}(X=1)) \,||\, P(Y|\mathrm{do}(X=0))\Big) \\ &= H\Big[P(X=1)P(Y|\mathrm{do}(X=1)) + P(X=0)P(Y|\mathrm{do}(X=0))\Big] \\ &\quad - P(X=1)H\big[P(Y|\mathrm{do}(X=1))\big] - P(X=0)H\big[P(Y|\mathrm{do}(X=0))\big] \\ &\overset{(1)}{=} H\Big[P(X=1)P(Y|X=1) + P(X=0)P(Y|X=0)\Big] \\ &\quad - P(X=1)H\big[P(Y|X=1)\big] - P(X=0)H\big[P(Y|X=0)\big] \\ &= H[P(Y)] - \mathbb{E}_X H[P(Y|X=x)] \\ &= I(X;Y), \end{aligned} \tag{A1}$$

where (1) holds since $I(Y \to X) = 0$ implies $P(Y|\mathrm{do}(X=x)) = P(Y|X=x)$ for all x, which follows from Proposition 1 and Definition 3. □

Proof of Proposition 4.

$$\begin{aligned} JSD(r_z \,||\, q_z) &= JSD\Big(P(Y|\mathrm{do}(X=1), Z=z) \,||\, P(Y|\mathrm{do}(X=0), Z=z)\Big) \\ &= H\Big[P(X=1|Z=z)P(Y|\mathrm{do}(X=1), Z=z) + P(X=0|Z=z)P(Y|\mathrm{do}(X=0), Z=z)\Big] \\ &\quad - P(X=1|Z=z)H\big[P(Y|\mathrm{do}(X=1), Z=z)\big] - P(X=0|Z=z)H\big[P(Y|\mathrm{do}(X=0), Z=z)\big] \\ &\overset{(1)}{=} H\Big[P(X=1|Z=z)P(Y|X=1, Z=z) + P(X=0|Z=z)P(Y|X=0, Z=z)\Big] \\ &\quad - P(X=1|Z=z)H\big[P(Y|X=1, Z=z)\big] - P(X=0|Z=z)H\big[P(Y|X=0, Z=z)\big] \\ &\overset{(2)}{=} H[P(Y|Z=z)] - H[P(Y|X, Z=z)] \\ &= D_{KL}\Big(P(X, Y|Z=z) || P(X|Z=z)P(Y|Z=z)\Big), \end{aligned} \tag{A2}$$

where (1) holds since $I(Y \to X|Z) = 0$ implies $P(Y|\mathrm{do}(X=x), Z=z) = P(Y|X=x, Z=z)$ for all x and z, which follows from Proposition 2 and Definition 2.

(2) holds since

$$H\Big[P(X=1|Z=z)P(Y|X=1, Z=z) + P(X=0|Z=z)P(Y|X=0, Z=z)\Big] = H[P(Y|Z=z)] \tag{A3}$$

and

$$\begin{aligned} &P(X=1|Z=z)H\big[P(Y|\mathrm{do}(X=1), Z=z)\big] + P(X=0|Z=z)H\big[P(Y|\mathrm{do}(X=0), Z=z)\big] \\ &\sum_x P(X=x|Z=z) \sum_y P(Y=y|X=x, Z=z) \log P(Y=y|X=x, Z=z) \\ &= \sum_{x,y} P(Y=y, X=x|Z=z) \log P(Y=y|X=x, Z=z) \\ &= H[P(Y|X, Z=z)]. \end{aligned} \tag{A4}$$

□

Proof of Corollary 1. By Proposition 4 we have:

$$\mathbb{E}_Z\left[JSD(r_z \,||\, q_z)\right] = \mathbb{E}_Z\Big[D_{KL}(P(X, Y|Z=z) \,||\, P(X|Z=z)P(Y|Z=z))\Big] \tag{A5}$$

and further

$$\begin{aligned}\mathbb{E}_Z\Big[D_{KL}(P(X,Y|Z=z)\;||\;P(X|Z=z)P(Y|Z=z))\Big]\\ = D_{KL}\Big(P(X,Y,Z)\;||\;P(X|Z)P(Y|Z)P(Z)\Big) = I(X;Y|Z).\end{aligned} \tag{A6}$$

□

Proofs of Theorems 2 to 4 all apply Definitions 6 and 7, expanded according to Equations (A7) and (A8), respectively. Subsequently, factorisations based on appropriate graphical models (Pearlian DAGs, chain graphs) are used.

$$I(X\to Y) = \mathbb{E}_{P(X,Y)}\log\left[\frac{P(X|Y)}{P(X|\mathrm{do}(Y))}\right] = \mathbb{E}_{P(X,Y)}\log\left[\frac{P(X,Y)}{P(X|\mathrm{do}(Y))P(Y)}\right] \tag{A7}$$

$$I(X\to Y|Z) = \mathbb{E}_{P(X,Y,Z)}\log\left[\frac{P(X|Y,Z)}{P(X|\mathrm{do}(Y),Z)}\right] = \mathbb{E}_{P(X,Y)}\log\left[\frac{P(X,Y,Z)}{\frac{P(X,Z|\mathrm{do}(Y))}{P(Z|\mathrm{do}(Y))}P(Y,Z)}\right] \tag{A8}$$

Proof of Theorem 2. Apply Definition 6 and Equation (A7) and factorise observational and interventional distributions P according to Pearlian DAG $\mathcal{G}_1$ from Figure 3 and Equations (1) and (3), respectively.

$$\begin{aligned}I(X^n\to Y^n) &= \mathbb{E}\log\frac{P(X^n,Y^n)}{P(X^n|\mathrm{do}(Y^n))P(Y^n)} = \mathbb{E}\log\frac{\prod_{i=1}^n P(Y_i|X^i,Y^{i-1})\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1})}{\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1})\prod_{i=1}^n P(Y_i|Y^{i-1})} =\\ &= \mathbb{E}\log\prod_{i=1}^n\frac{P(X^i,Y^i)}{P(X^i,Y^{i-1})P(Y_i|Y^{i-1})} = \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^i,Y^i)}{P(Y^{i-1})}}{\frac{P(X^i,Y^{i-1})}{P(Y^{i-1})}P(Y_i|Y^{i-1})} =\\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(X^i,Y_i|Y^{i-1})}{\prod_{i=1}^n P(X^i|Y^{i-1})P(Y_i|Y^{i-1})} = \sum_{i=1}^n I(X^i;Y_i|Y^{i-1}),\end{aligned} \tag{A9}$$

$$\begin{aligned}I(Y^n\to X^n) &= \mathbb{E}\log\frac{P(X^n,Y^n)}{P(Y^n|\mathrm{do}(X^n))P(X^n)} = \mathbb{E}\log\frac{\prod_{i=1}^n P(Y_i|X^i,Y^{i-1})\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1})}{\prod_{i=1}^n P(Y_i|X^i,Y^{i-1})\prod_{i=1}^n P(X_i|X^{i-1})} =\\ &= \mathbb{E}\log\prod_{i=1}^n\frac{P(X^i,Y^{i-1})}{P(X^{i-1},Y^{i-1})P(X_i|X^{i-1})} = \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^i,Y^{i-1})}{P(X^{i-1})}}{\frac{P(X^{i-1},Y^{i-1})}{P(X^{i-1})}P(X_i|X^{i-1})} =\\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(Y^{i-1},X_i|X^{i-1})}{\prod_{i=1}^n P(Y^{i-1}|X^{i-1})P(X_i|X^{i-1})} = \sum_{i=1}^n I(Y^{i-1};X_i|X^{i-1}),\end{aligned} \tag{A10}$$

All expectations are taken with respect to $P(X,Y)$. □

Proof of Theorem 3. Apply Definition 6 and Equation (A7) and factorise observational and interventional distributions P according to Pearlian DAG $\mathcal{G}_2$ from Figure 4a and Equations (1) and (3), respectively.

$$\begin{aligned} I(X^n \to Y^n) &= \mathbb{E}\log\frac{P(X^n,Y^n)}{P(X^n|\mathrm{do}(Y^n))P(Y^n)} = \mathbb{E}\log\frac{\prod_{i=1}^n P(Y_i|X^{i-1},Y^{i-1})\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1})}{\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1})\prod_{i=1}^n P(Y_i|Y^{i-1})} = \\ &= \mathbb{E}\log\prod_{i=1}^n\frac{P(X^{i-1},Y^i)}{P(X^{i-1},Y^{i-1})P(Y_i|Y^{i-1})} = \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^{i-1},Y^i)}{P(Y^{i-1})}}{\frac{P(X^{i-1},Y^{i-1})}{P(Y^{i-1})}P(Y_i|Y^{i-1})} = \\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(X^{i-1},Y_i|Y^{i-1})}{\prod_{i=1}^n P(X^{i-1}|Y^{i-1})P(Y_i|Y^{i-1})} = \sum_{i=1}^n I(X^{i-1};Y_i|Y^{i-1}), \end{aligned} \tag{A11}$$

All expectations are taken with respect to $P(X,Y)$.

Equation (21) is analogous to Equation (20). We now consider the Pearlian DAG from Figure 4b and Definition 7 and Equation (A8) instead of Definition 6 and Equation (A7).

$$\begin{aligned} I(X^n \to Y^n|Z^n) &= \mathbb{E}\log\frac{P(X^n,Y^n,Z^n)}{P(X^n|\mathrm{do}(Y^n),Z^n)P(Y^n,Z^n)} = \mathbb{E}\log\frac{P(X^n,Y^n,Z^n)}{\frac{P(X^n,Z^n|\mathrm{do}(Y^n))}{P(Z^n|\mathrm{do}(Y^n))}P(Y^n,Z^n)} = \\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(Y_i|X^{i-1},Y^{i-1},Z^{i-1})\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1},Z^{i-1})\prod_{i=1}^n P(Z_i|X^{i-1},Y^{i-1},Z^{i-1})}{\frac{\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1},Z^{i-1})\prod_{i=1}^n P(Z_i|X^{i-1},Y^{i-1},Z^{i-1})}{\prod_{i=1}^n P(Z_i|Y^{i-1},Z^{i-1})}\prod_{i=1}^n P(Y_i|Y^{i-1},Z^{i-1})\prod_{i=1}^n P(Z_i|Y^{i-1},Z^{i-1})} = \\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(Y_i|X^{i-1},Y^{i-1},Z^{i-1})}{\prod_{i=1}^n P(Y_i|Y^{i-1},Z^{i-1})} = \mathbb{E}\log\prod_{i=1}^n\frac{P(X^{i-1},Y^i,Z^{i-1})}{P(X^{i-1},Y^{i-1},Z^{i-1})P(Y_i|Y^{i-1},Z^{i-1})} = \\ &= \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^{i-1},Y^i,Z^{i-1})}{P(Y^{i-1},Z^{i-1})}}{\frac{P(X^{i-1},Y^{i-1},Z^{i-1})}{P(Y^{i-1},Z^{i-1})}P(Y_i|Y^{i-1},Z^{i-1})} = \\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(X^{i-1},Y_i|Y^{i-1},Z^{i-1})}{\prod_{i=1}^n P(X^{i-1}|Y^{i-1},Z^{i-1})P(Y_i|Y^{i-1},Z^{i-1})} = \sum_{i=1}^n I(X^{i-1};Y_i|Y^{i-1},Z^{i-1}), \end{aligned} \tag{A12}$$

here, all expectations are taken with respect to $P(X,Y,Z)$. □

We now switch the underlying graphical model from Pearlian DAGs to chain graphs. Note that the difference to the proof of Theorem 3 lies in the inclusion of chain components in the factorisation of observational and interventional distributions.

Proof of Theorem 4. Apply Definition 6 and Equation (A7) and factorise observational and interventional distributions P according to the chain graph $\mathcal{H}$ from Figure 6a and Equations (22) and (23), respectively.

$$\begin{aligned} I(X^n \to Y^n) &= \mathbb{E}\log\frac{P(X^n,Y^n)}{P(X^n|\mathrm{do}(Y^n))P(Y^n)} = \mathbb{E}\log\frac{\prod_{i=1}^n P(X_i,Y_i|X^{i-1},Y^{i-1})}{\prod_{i=1}^n P(X_i|X^{i-1},Y^{i-1},Y_i)\prod_{i=1}^n P(Y_i|Y^{i-1})} = \\ &= \mathbb{E}\log\prod_{i=1}^n\frac{P(X^i,Y^i)P(X^{i-1},Y^i)}{P(X^{i-1},Y^{i-1})P(X^i,Y^i)P(Y_i|Y^{i-1})} = \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^{i-1},Y^i)}{P(Y^{i-1})}}{\frac{P(X^{i-1},Y^{i-1})}{P(Y^{i-1})}P(Y_i|Y^{i-1})} = \\ &= \mathbb{E}\log\frac{\prod_{i=1}^n P(X^{i-1},Y_i|Y^{i-1})}{\prod_{i=1}^n P(X^{i-1}|Y^{i-1})P(Y_i|Y^{i-1})} = \sum_{i=1}^n I(X^{i-1};Y_i|Y^{i-1}), \end{aligned} \tag{A13}$$

All expectations are taken with respect to $P(X,Y)$.

Equation (24) is analogous to Equation (25). We now consider the chain graph from Figure 6b and Definition 7 and Equation (A8) instead of Definition 6 and Equation (A7).

$$\begin{aligned}
I(X^n \to Y^n|Z^n) &= \mathbb{E}\log\frac{P(X^n,Y^n,Z^n)}{P(X^n|\mathrm{do}(Y^n),Z^n)P(Y^n,Z^n)} = \mathbb{E}\log\frac{P(X^n,Y^n,Z^n)}{\frac{P(X^n,Z^n|\mathrm{do}(Y^n))}{P(Z^n|\mathrm{do}(Y^n))}P(Y^n,Z^n)} = \\
&= \mathbb{E}\log\frac{\prod_{i=1}^n P(X_i,Y_i,Z_i|X^{i-1},Y^{i-1},Z^{i-1})}{\frac{\prod_{i=1}^n P(X_i,Z_i|X^{i-1},Y^{i-1},Z^{i-1},Y_i)}{\prod_{i=1}^n P(Z_i|Y^{i-1},Z^{i-1},Y_i)}\prod_{i=1}^n P(Y_i,Z_i|Y^{i-1},Z^{i-1})} = \\
&= \mathbb{E}\log\prod_{i=1}^n\frac{P(X^i,Y^i,Z^i)}{P(X^{i-1},Y^{i-1},Z^{i-1})}\frac{P(X^{i-1},Y^i,Z^{i-1})}{P(X^i,Y^i,Z^i)}\frac{P(Y^i,Z^i)}{P(Y^i,Z^{i-1})}\frac{P(Y^{i-1},Z^{i-1})}{P(Y^i,Z^i)} = \\
&= \mathbb{E}\log\prod_{i=1}^n\frac{\frac{P(X^{i-1},Y^i,Z^{i-1})}{P(Y^{i-1},Z^{i-1})}}{\frac{P(X^{i-1},Y^{i-1},Z^{i-1})}{P(Y^{i-1},Z^{i-1})}P(Y_i|Y^{i-1},Z^{i-1})} = \\
&= \mathbb{E}\log\frac{\prod_{i=1}^n P(X^{i-1},Y_i|Y^{i-1},Z^{i-1})}{\prod_{i=1}^n P(X^{i-1}|Y^{i-1},Z^{i-1})P(Y_i|Y^{i-1},Z^{i-1})} = \sum_{i=1}^n I(X^{i-1};Y_i|Y^{i-1},Z^{i-1}),
\end{aligned} \tag{A14}$$

here, all expectations are taken with respect to $P(X,Y,Z)$. □

References

1. Clarke, B. Causality in Medicine with Particular Reference to the Viral Causation of Cancers. Ph.D. Thesis, University College London, London, UK, January 2011.
2. Rasmussen, S.A.; Jamieson, D.J.; Honein, M.A.; Petersen, L.R. Zika virus and birth defects—Reviewing the evidence for causality. *N. Engl. J. Med.* **2016**, *374*, 1981–1987.
3. Samarasinghe, S.; McGraw, M.; Barnes, E.; Ebert-Uphoff, I. A study of links between the Arctic and the midlatitude jet stream using Granger and Pearl causality. *Environmetrics* **2019**, *30*, e2540.
4. Dourado, J.R.; Júnior, J.N.d.O.; Maciel, C.D. Parallelism Strategies for Big Data Delayed Transfer Entropy Evaluation. *Algorithms* **2019**, *12*, 190.
5. Peia, O.; Roszbach, K. Finance and growth: Time series evidence on causality. *J. Financ. Stabil.* **2015**, *19*, 105–118.
6. Soytas, U.; Sari, R. Energy consumption and GDP: Causality relationship in G-7 countries and emerging markets. *Energy Econ.* **2003**, *25*, 33–37.
7. Dippel, C.; Gold, R.; Heblich, S.; Pinto, R. Instrumental Variables and Causal Mechanisms: Unpacking the Effect of Trade on Workers and Voters. Technical Report. National Bureau of Economic Research. 2017. Available online: https://www.nber.org/papers/w23209 (accessed on 2 October 2019).
8. Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; Peters, J. Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **2018**, *19*, 1309–1342.
9. Spirtes, P.; Glymour, C.N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; Richardson, T. *Causation, Prediction, and Search*; MIT Press: Cambridge, MA, USA, 2000.
10. Verma, T.; Pearl, J. Equivalence and Synthesis of Causal Models. In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 27–29 July 1990; Elsevier Science Inc.: New York, NY, USA, 1991; pp. 255–270.
11. Massey, J.L. Causality, feedback and directed information. In Proceedings of the International Symposium on Information Theory and Its Applications, Waikiki, HI, USA, 27–30 November 1990.
12. Eichler, M. Graphical modelling of multivariate time series. *Probab. Theory Relat. Fields* **2012**, *153*, 233–268. doi:10.1007/s00440-011-0345-8.
13. Quinn, C.J.; Kiyavash, N.; Coleman, T.P. Directed information graphs. *IEEE Trans. Inf. Theory* **2015**, *61*, 6887–6909.
14. Tatikonda, S.; Mitter, S. The capacity of channels with feedback. *IEEE Trans. Inf. Theory* **2009**, *55*, 323–349.

15. Raginsky, M. Directed information and Pearl's causal calculus. In Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 28–30 Septemper 2011; pp. 958–965.
16. Marko, H. The Bidirectional Communication Theory-A Generalization of Information Theory. *IEEE Trans. Commun.* **1973**, *21*, 1345–1351. doi:10.1109/TCOM.1973.1091610.
17. Granger, C. Economic processes involving feedback. *Inf. Control* **1963**, *6*, 28–48.
18. Granger, C. Testing for causality: A personal viewpoint. *J. Econ. Dyn. Control* **1980**, *2*, 329–352.
19. Kramer, G. Directed Information for Channels with Feedback. Ph.D. Thesis, ETH Zurich, Zürich, Switzerland, 1998.
20. Amblard, P.O.; Michel, O.J.J. The Relation between Granger Causality and Directed Information Theory: A Review. *Entropy* **2013**, *15*, 113–143. doi:10.3390/e15010113.
21. Amblard, P.O.; Michel, O. Causal Conditioning and Instantaneous Coupling in Causality Graphs. *Inf. Sci.* **2014**, *264*, 279–290. doi:10.1016/j.ins.2013.12.037.
22. Quinn, C.J.; Coleman, T.P.; Kiyavash, N. Causal dependence tree approximations of joint distributions for multiple random processes. *arXiv* **2011**, arXiv:1101.5108.
23. Quinn, C.J.; Kiyavash, N.; Coleman, T.P. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Trans. Signal Process.* **2013**, *61*, 3173–3182.
24. Weissman, T.; Kim, Y.; Permuter, H.H. Directed Information, Causal Estimation, and Communication in Continuous Time. *IEEE Trans. Inf. Theory* **2013**, *59*, 1271–1287. doi:10.1109/TIT.2012.2227677.
25. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
26. Eichler, M. Causal inference with multiple time series: principles and problems. *Philos. Trans. R. Soc. A* **2013**, *371*, 20110613.
27. Jafari-Mamaghani, M.; Tyrcha, J. Transfer entropy expressions for a class of non-Gaussian distributions. *Entropy* **2014**, *16*, 1743–1755.
28. Ay, N.; Polani, D. Information flows in causal networks. *Adv. Complex Syst.* **2008**, *11*, 17–41.
29. Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
30. James, R.G.; Barnett, N.; Crutchfield, J.P. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.* **2016**, *116*, 238701.
31. Sharma, A.; Sharma, M.; Rhinehart, N.; Kitani, K.M. Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using Directed Information. *arXiv* **2018**, arXiv:1810.01266.
32. Tanaka, T.; Skoglund, M.; Sandberg, H.; Johansson, K.H. Directed information and privacy loss in cloud-based control. In Proceedings of the 2017 American Control Conference (ACC), Seattle, WA, USA, 24–26 May 2017; pp. 1666–1672.
33. Tanaka, T.; Esfahani, P.M.; Mitter, S.K. LQG control with minimum directed information: Semidefinite programming approach. *IEEE Trans. Autom. Control* **2018**, *63*, 37–52.
34. Etesami, J.; Kiyavash, N.; Coleman, T. Learning Minimal Latent Directed Information Polytrees. *Neural Comput.* **2016**, *28*, 1723–1768. PMID: 27391682, doi:10.1162/NECO_a_00874.
35. Zhou, Y.; Spanos, C.J. Causal meets Submodular: Subset Selection with Directed Information. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 2649–2657.
36. Mehta, K.; Kliewer, J. Directional and Causal Information Flow in EEG for Assessing Perceived Audio Quality. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2017**, *3*, 150–165.
37. Zaremba, A.; Aste, T. Measures of causality in complex datasets with application to financial data. *Entropy* **2014**, *16*, 2309–2349.
38. Diks, C.; Fang, H. Transfer Entropy for Nonparametric Granger Causality Detection: An Evaluation of Different Resampling Methods. *Entropy* **2017**, *19*, 372.
39. Soltani, N.; Goldsmith, A.J. Directed information between connected leaky integrate-and-fire neurons. *IEEE Trans. Inf. Theory* **2017**, *63*, 5954–5967.
40. Kontoyiannis, I.; Skoularidou, M. Estimating the Directed Information and Testing for Causality. *IEEE Trans. Inf. Theory* **2016**, *62*, 6053–6067. doi:10.1109/TIT.2016.2604842.

41. Charalambous, C.D.; Stavrou, P.A. Directed information on abstract spaces: Properties and variational equalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 6019–6052.
42. Lauritzen, S.L. *Graphical Models*; Clarendon Press: Oxford, UK, 1996; Volume 17.
43. Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M.H.; Bühlmann, P.; others. Causal inference using graphical models with the R package pcalg. **2012**, doi:10.18637/jss.v047.i11.
44. Richardson, T.; Spirtes, P. Ancestral graph Markov models. *Ann. Stat.* **2002**, *30*, 962–1030.
45. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **2008**, *172*, 1873–1896.
46. Pearl, J. Causal diagrams for empirical research. *Biometrika* **1995**, *82*, 669–688.
47. Pearl, J. *The Causal Foundations of Structural Equation Modeling*. Technical Report; DTIC Document; Guilford Press: New York, NY, USA, 2012.
48. Lauritzen, S.L.; Wermuth, N. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Stat.* **1989**; pp. 31–57.
49. Sonntag, D. A Study of Chain Graph Interpretations. Ph.D. Thesis, Linköping University, Linköping, Sweden, 2014.
50. Lauritzen, S.L.; Wermuth, N. *Mixed Interaction Models*; Institut for Elektroniske Systemer: Aalborg Universitetscenter, Aalborg, Denmark, 1984.
51. Frydenberg, M. The chain graph Markov property. *Scand. J. Stat.* **1990**, *17*, 333–353.
52. Lauritzen, S.L.; Richardson, T.S. Chain graph models and their causal interpretations. *J. R. Stat. Soc. B* **2002**, *64*, 321–348.
53. Ogburn, E.L.; Shpitser, I.; Lee, Y. Causal inference, social networks, and chain graphs. *arXiv* **2018**, arXiv:1812.04990.
54. Andersson, S.A.; Madigan, D.; Perlman, M.D. Alternative Markov properties for chain graphs. *Scand. J. Stat.* **2001**, *28*, 33–85.
55. Cox, D.R.; Wermuth, N. *Multivariate Dependencies: Models, Analysis and Interpretation*; Chapman and Hall/CRC: London, UK, 2014.
56. Richardson, T. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **2003**, *30*, 145–157.
57. Peña, J.M. Alternative Markov and causal properties for Acyclic Directed Mixed Graphs. In Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, Jersey City, NJ, USA, 25–29 June 2016; pp. 577–586.
58. Peña, J.M. Learning acyclic directed mixed graphs from observations and interventions. In Proceedings of the Eighth International Conference on Probabilistic Graphical Models, Lugano, Switzerland, 6–9 September 2016; pp. 392–402.
59. Studenỳ, M. Bayesian networks from the point of view of chain graphs. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, USA, 24–26 July 1998; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1998; pp. 496–503.
60. Richardson, T.S. A Factorization Criterion for Acyclic Directed Mixed Graphs. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June, 2009; AUAI Press: Arlington, VA, USA, 2009; pp. 462–470.
61. Dawid, A.P. Beware of the DAG! In Proceedings of Workshop on Causality: Objectives and Assessment, Whistler, BC, Canada, 12 December 2008; MIT Press: Cambridge, MA, USA, 2010; Volume 6, pp. 59–86.
62. Pearl, J. An introduction to causal inference. *Int. J. Biostat.* **2010**, *6*, doi:10.2202/1557-4679.1203.
63. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **2009**, *3*, 96–146.
64. Rubin, D.B. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* **1978**, *6*, 34–58.
65. Spława-Neyman, J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **1923**, *10*, 1–51.
66. Spława-Neyman, J.; Dąbrowska, D.M.; Speed, T. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* **1990**, *5*, 465–472.
67. Imbens, G.W.; Rubin, D.B. *Causal Inference in Statistics, Social, and Biomedical Sciences*; Cambridge University Press: Cambridge, UK, 2015.
68. Dawid, A.P. Statistical causality from a decision-theoretic perspective. *Ann. Rev. Stat. Appl.* **2015**, *2*, 273–303.

69. Shpitser, I.; VanderWeele, T.; Robins, J.M. On the validity of covariate adjustment for estimating causal effects. In Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, 8–11 July 2010; AUAI Press: Arlington, VA, USA, 2010; pp. 527–536.
70. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.
71. Holland, P.W. Causal inference, path analysis and recursive structural equations models. *Sociol. Methodol.* **1988**, *8*, 449-484
72. Dawid, A.P. Fundamentals of Statistical Causality. Research Report No. 279. Available online: https://pdfs.semanticscholar.org/c4bc/ad0bb58091ecf9204ddb5db7dce749b0d461.pdf (accessed on 2 October 2019).
73. Guo, H.; Dawid, P. Sufficient covariates and linear propensity analysis. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 3–15 May 2010; Volume 9, pp. 281–288.
74. Imbens, G.W.; Wooldridge, J.M. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **2009**, *47*, 5–86.
75. Kallus, N.; Mao, X.; Zhou, A. Interval Estimation of Individual-Level Causal Effects Under Unobserved Confounding. In Procedings of the 22nd International Conference on Artificial Intelligence and Statistics, Okinawa, Japan, 16–18 April 2019; pp. 2281–2290.
76. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
77. Nielsen, F. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy* **2019**, *21*, 485.
78. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
79. DeDeo, S.; Hawkins, R.; Klingenstein, S.; Hitchcock, T. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy* **2013**, *15*, 2246–2276.
80. Contreras-Reyes, J.E. Analyzing fish condition factor index through skew-gaussian information theory quantifiers. *Fluctuation Noise Lett.* **2016**, *15*, 1650013.
81. Zhou, K.; Varadarajan, K.M.; Zillich, M.; Vincze, M. Gaussian-weighted Jensen–Shannon divergence as a robust fitness function for multi-model fitting. *Mach. Vis. Appl.* **2013**, *24*, 1107–1119. doi:10.1007/s00138-013-0513-1.
82. Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; Schölkopf, B. Quantifying causal influences. *Ann. Stat.* **2013**, *41*, 2324–2358.
83. Geiger, P.; Janzing, D.; Schölkopf, B. Estimating Causal Effects by Bounding Confounding. In Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, Quebec City, QC, Canada, 23–27 July 2014; AUAI Press: Arlington, VA, USA, 2014; pp. 240–249.
84. Sun, J.; Bollt, E.M. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D* **2014**, *267*, 49–57.
85. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
86. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv* **2014**, arXiv:1401.4082.
87. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep variational information bottleneck. *arXiv* **2016**, arXiv:1612.00410.
88. Wieczorek, A.; Wieser, M.; Murezzan, D.; Roth, V. Learning Sparse Latent Representations with the Deep Copula Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
89. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 2172–2180.
90. Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv* **2019**, arXiv:1901.10912.

91. Suter, R.; Miladinovic, D.; Schölkopf, B.; Bauer, S. Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6056–6065.
92. Besserve, M.; Sun, R.; Schölkopf, B. Counterfactuals uncover the modular structure of deep generative models. *arXiv* **2018**, arXiv:1812.03253.
93. Chattopadhyay, A.; Manupriya, P.; Sarkar, A.; Balasubramanian, V.N. Neural Network Attributions: A Causal Perspective. *arXiv* **2019**, arXiv:1902.02302.