entropy

MDPI

Article

Relaxation of Some Confusions about Confounders

Ádám Zlatniczki [1,2], Marcell Stippinger [3], Zsigmond Benkő [3], Zoltán Somogyvári [3] and András Telcs [1,2,3,4,*]

1 Department of Computer Science and Information Theory, Budapest University of Technology and Economics, H-1111 Budapest, Hungary; adam.zlatniczki@ericsson.com
2 Ericsson Hungary, H-1117 Budapest, Hungary
3 Wigner Research Centre for Physics, H-1121 Budapest, Hungary; stippinger.marcell@wigner.hu (M.S.); benko.zsigmond@wigner.hu (Z.B.); somogyvari.zoltan@wigner.hu (Z.S.)
4 Department of Quantitative Methods, University of Pannonia, H-8200 Veszprém, Hungary
* Correspondence: telcs.andras@wigner.hu; Tel.: +36-30-375-3896

Abstract: This work is about observational causal discovery for deterministic and stochastic dynamic systems. We explore what additional knowledge can be gained by the usage of standard conditional independence tests and if the interacting systems are located in a geodesic space.

Keywords: causality; common cause; geodesic

Citation: Zlatniczki, Á.; Stippinger, M.; Benkő, Z.; Somogyvári, Z.; Telcs, A. Relaxation of Some Confusions about Confounders. *Entropy* **2021**, *23*, 1450. https://doi.org/10.3390/e23111450

Academic Editor: Ivanka Stamova

Received: 2 October 2021
Accepted: 26 October 2021
Published: 31 October 2021

1. Introduction

It is not necessary to emphasize the importance of the concept of causality in science and in the natural sciences in particular. The concept traverses all disciplines, and it is a matter of extensive research fueled by the exponentially increasing available scientific data and computation power. Revealing causal relations between systems via the time series produced by them is one of the most attractive challenges. The first major advancement was due to Granger who used an auto-regressive framework for a practical implementation of the predictive causality principle by Wiener [1].

The very popular Granger [2] method has some theoretical and practical limitations. It is not able to detect hidden common cause and, instead, indicates false directional causal relation between the observed systems (for details of all the pros and cons cf. [3]). Several further methods appeared in the last two decades (for a concise review see Runge [4] or [5,6]). One of the most prominent is the convergent cross mapping method developed by Sugihara [7] to investigate deterministic dynamic systems, which essentially utilizes Takens' embedding theorem [8]. Stark [9,10] generalized Takens' result and showed the theoretical limitations to use it for stochastic dynamic systems. For deterministic dynamics, a new approach was presented in a recent work [11] that was based on the comparison of the dimension of the attractors of the given systems and their joint observation.

The present paper investigates the causal relation of a pair of dynamic systems (which might be deterministic or stochastic). Facts are revealed that, to our best knowledge, avoided the attention of previous studies. We show that the common driver is an i.i.d. sequence, shared observational noise, if there is dependence between the systems with the smallest but positive time difference. We also show that, if the pair is located in a non-abstract physical space where the speed of information transfer is known, then direct causation and common cause cases can be distinguished, which, in general, is theoretically impossible.

Basic Definitions

First, we provide the framework of our investigation. Our aim is to find the causal relationship between two stochastic dynamic systems X and Y from which we observe the time series $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$.

Assumption 1. *We assume that there is a set of systems* $S = \{X, Y, L\}, X \in \mathbb{R}^{d_X}$ *and* $Y \in \mathbb{R}^{d_Y}, L \in \mathbb{R}^{d_L}$, $D = d_X + d_Y + d_L$, *and an external source of noise* W *such that the process* $m_i = (s_i, \omega_i) = \left((x_i, y_i, l_i^1, \dots l_i^m), \omega_i\right) \in \mathbb{R}^{2D}, s_i \in \mathbb{R}^D, \omega_i \in \mathbb{R}^D$ *has a joint distribution. The series* l_n *are unobserved, hidden series that are not i.i.d. (see Figure 1).*

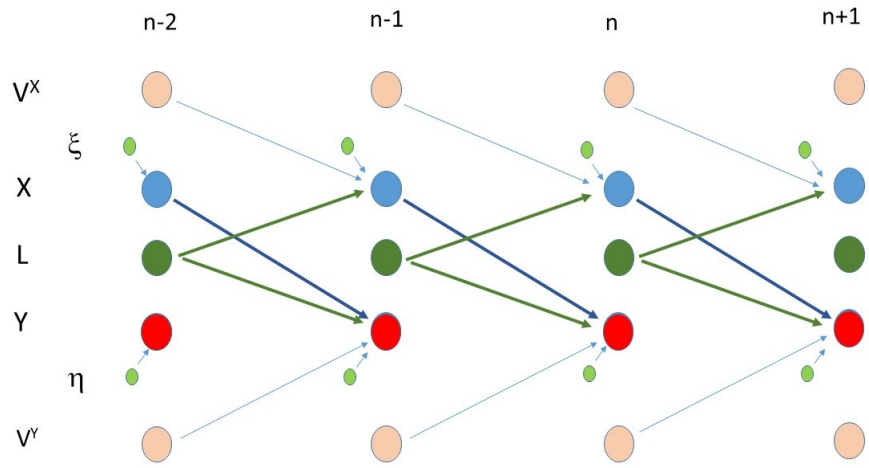

Figure 1. An example is given for a possible causation scheme for the system M. In the observed series X and Y, and in between we have L a common cause. Above X and below Y, small circles represent the i.i.d. input ξ, η, and the large circles V^X, V^Y (also belonging to the set of unobserved series) represent the non i.i.d. influences that are not shared and not common for X and Y. In this example, X drives Y, and they have L as a common cause.

In what follows, for ω_1, we will use ξ emphasizing that it influences x and similarly η for ω_2 for y.

For brevity, we will use the multi-index of involved dimensions: $\underline{d} = (d_X, d_Y, d_L)$.

Assumption 2. *The external noise* $\omega_i \in \mathbb{R}^D$ *is modeled with an unobserved i.i.d. sequence and affects all the systems with independent* ξ, η, ω_{l_i} *components. Furthermore,* ω_{i+1} *is independent from* $S_0^i = \{s_j\}_{j=0}^i$.

Assumption 3. *The process* m_i *is stationary.*

Assumption 4. *The causal structure of the time series is time invariant and non random.*

In what follows, we use the expression "drives" for all the terms "causes", "influences" and "injects information" in relation to dynamic systems.

Following [12], we use the next model. The visible and invisible system can be described by a p-order Structural Vector Auto-regressive $SVAR$ model:

$$m_{n+1} = f\left(m_n, \dots, m_{n-p+1}, \omega_{n+1}\right)$$

$n \in \mathbb{N}$, and m_0 follows the stationary distribution of the system. (It is a $SVAR(\underline{d}, p)$ process, were $\underline{d}$ is the multi index of dimensions of the variables, and p is the order of auto-regression.)

The recursion clearly can be transformed with time delay embedding into higher dimension first order $SVAR$ in particular to $SVAR(2D \times p, 1)$ in short $SVAR(1)$ with

$$M_n = \left(m_n, m_{n-1}, \dots, m_{n-p+1}\right) : \tag{1}$$

$$M_{n+1} = g(M_n, \omega_{n+1}) \tag{2}$$

We make the same restriction as in [12] that (2) must be recursive in the variables, which ensures that there is no directed functional cycle. Variables with capital letters denote the same "embedding" as in (1).

Assumption 5. *The process M is exact (cf. Definition 4.3.2. [13]).*

Exactness means that, if the process started from a set with positive probability, then, after a long time, the set in which it can be found has probability one. It is natural to assume exactness given that we work with an observation, and the support of the observed process for us will be the whole set where the process can run and, consequently, has probability one. On the other hand, exactness implies mixing for stationary processes, and, at the same time, a mixing stationary process is α-mixing (or strong mixing). Let us note here that, from strong mixing, ergodicity also follows, but we do not need that fact (see also [13]).

In our discovery scheme, we may allow instantaneous causation between all variables; however, we do not elaborate on that case here. For brevity, that is not reflected in (2). We note that a system like (2) with contemporaneous interaction but without a directed cycle can be rewritten into the form of (2) using time shifts thanks to the acyclic recursivity.

Definition 1. *We will say that X drives Y if there is a $k > 0$ s.t.*

$$Y_{n+k} = f(X_n, Y_n, L_n, \eta_{n+k}) \tag{3}$$

where L_n stands for the set of latent variables, η_{n+1} is an i.i.d sequence that is independent of $(X_i, Y_i, L_i)_{i=0}^{n}$, and X_n cannot be omitted without violating the validity of (3).

Let us explain that key definition. We may say that there is no such a function g that

$$Y_{n+k} = g(Y_n, L_n, \eta_{n+k}) \tag{4}$$

which makes the fact explicit, that Y_{n+1} can be created without X_n. Here, one should also observe that the i.i.d. part is also the same as in (3), and there is no possibility for an i.i.d. X_n to be hidden in η_{n+1}.

2. Causal Discovery Schemes

The literature of causal discovery is huge. This work has been inspired by two recent ones with their strengths and limitations. First, we found the framework defined by Malinsky in [12] very appealing and that the complex nature of assumptions and the suggested algorithm in [14] presented an essential challenge. The algorithm in [12] is an extension of [15,16]. The algorithm provides a theoretically complete recall of the underlying causal structure at the price that some relations are marked undetermined and some causal relations are not or only partially revealed.

In [14], in addition to many other assumptions, it is assumed that all hidden processes that influence an observed one have no memory (Assumption A9 in [14]). That assumption and A6 in [14] cannot be checked. In [12], such restrictions are eliminated. That paper and most of the works based on Pearl's DAG analysis have theoretical limitations as admitted in [12]. In what follows, we investigate some situations in which that limitation can be relaxed.

Information from X to Y can be transferred along a chain of direct causal links, along a directed path $\pi_{X,Y}$. The length of the path (the number of intermediate components plus one) is denoted by $l = l_{X,Y} = l(\pi_{X,Y})$. Such a path has a starting and ending time n, $n + l$ (for arbitrary $n \geq 0, l > 0$), the difference is the time lag.

Assumption 6. *We assume that, with some background information, the minimal lag between the systems X and Y can be determined.*

We consider the smallest lag τ for which dependence can be detected in "direction" X to Y:

$$\tau = \tau_{X,Y} = \min_{\pi}\{l(\pi_{X,Y}) > 0\} \tag{5}$$

2.1. The Decomposable Case

We introduce our notation. In order to save space, let $(A, B) = (X, Y)$ or (Y, X). Let I stand for the Shannon entropy/differential entropy based mutual information. We define conditional mutual information between elements of time series a_n, b_n and similarly for other series. A segment from k to l of a time series a_n are denoted by A_k^l. Such segments are used in the condition representing a part or the full past. In order to investigate if there is information transfer from B to A with a given time lag $\tau_{b,a}$ we use the conditional mutual information between $a_{n+\tau_{b,a}}$ and b_n given the full past of both series ${A_0}^{n+\tau_{b,a}-1}$ and B_0^{n-1}, and we denote it by I_B. We define the following conditional mutual information

$$\begin{aligned} I_B &= I\left(a_{n+\tau_{b,a}}; b_n | {A_0}^{n+\tau_{b,a1}-1}, B_0^{n-1}\right), \\ I_{A,B}^{(k)} &= I\left(a_n; b_{n+k} | A_0^{n-1}, B_0^{n+k-1}\right), \text{for any } 0 \le k < \tau_{A,B}. \end{aligned}$$

where we set $A_p^q = \left(a_q, \ldots, a_p\right)$ and similarly for B and other variables.

Proposition 1. *Let for L, A, B ($(A, B) = (X, Y)$ or (Y, X))*

$$\delta = \delta_{L,A,B} = \tau_{L,B} - \tau_{L,A} \ge 0.$$

Under Assumptions 1–6 for $\delta_{L,A,B} = k, 0 \le k < \tau_{A,B}$, the following implications hold.

$$\left\{ \begin{array}{cccc} I_A & I_B & & A \to B \quad B \to A \\ = c' & = 0 & \Leftrightarrow & \exists \quad \nexists \\ = c' & = c'' & \Leftrightarrow & \exists \quad \exists \end{array} \right\} \times \left\{ \begin{array}{ccc} I_{A,B}^{(k)} & & CD \\ = 0 & \Leftrightarrow & \nexists \\ = c & \Leftrightarrow & \exists \end{array} \right\},$$

Relation 1. Logical relations between conditional mutual information values and causal relations

where CD stands for Common Driver and $c, c', c'' > 0$. In the right part of the table, =0 means that $I_{A,B}^{(k)} = 0$ holds for all $0 \le k < \tau_{A,B}$, while >0 means that there is at least one such k for which $I_{A,B}^{(k)} > 0$.

The proposition summarizes the possible inferences in a concise way. In Relation 1, the headers contain the list of possible combinations and the possible causal scenarios. We have the direct product of two lists of cases collected in the two tables. The header of tables contains, on the left, the quantities that are decisive and, on the right, the possible causal scenarios. As an example, in the left table, the first row shows that if and only if we have that $I_B = 0$ but $I_A = c' > 0$ (significantly differ from zero) then B does not drive A but A drives B. In the right table, if $I_{A,B}^{(k)} = 0$ holds for all $0 \le k < \tau_{a,b}$ that means that there is no common information between members of the series for $k < \tau_{a,b}$, while, in the opposite case, there should be a common driver, given, that there is shared information that cannot be attributed to driving with a lag below $\tau_{a,b}$. If $\delta = \tau_{x,y}$ (or $= \tau_{y,x}$) then, causation between X and Y and a common driver may coexist, and we cannot separate those models. In the next section we provide some observations in that situation.

2.2. The Confounder Case

We assume that $\delta_{L,X,Y} = \tau_{X,Y}$ but $\tau > 0$. If $\tau > 0$, we can investigate the common information between X_{n+1} and $Y_{n+\tau}$. Unfortunately, the variables $X_n, Y_{n+\tau}$ have a confounder; therefore, we cannot tell which causal relation is behind the dependence. However, some

internal structure can be revealed. In line with the assumptions $\delta_{L,X,Y} = \tau_{X,Y}$ but $\tau > 0$, we assume that

$$I_0 = I\left(X_n; Y_{n+\tau} | X_0^{n-1}, Y_0^{n+\tau-1}\right) > 0 \quad (6)$$

$$I_1 = I\left(X_{n+1}; Y_{n+\tau} | X_0^{n}, Y_0^{n+\tau-1}\right) = 0 \quad (7)$$

Let b_1 be the information that is passed from X_n to $Y_{n+\tau}$ and b_i for $i = 1, 2$ from an L to both (if one or other information transfer takes place). We also let a_1 be the information passed from X_n to the X_{n+1} as Figure 2 shows.

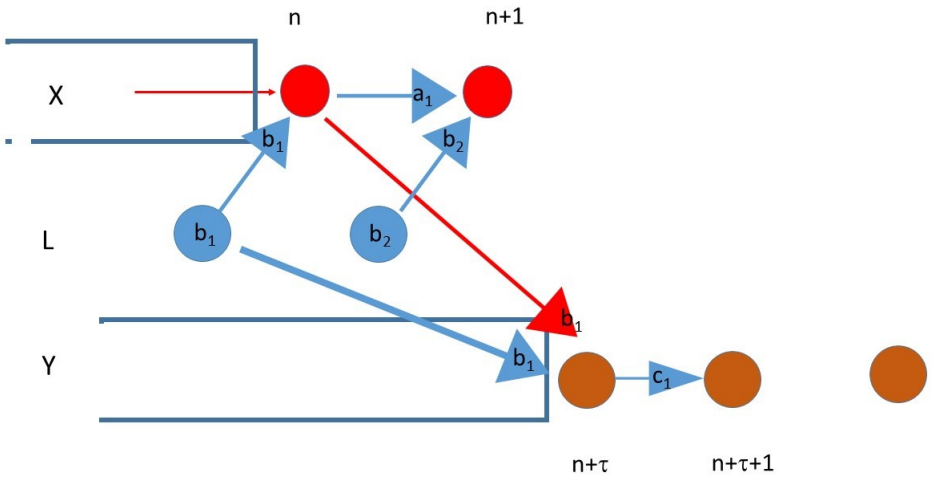

Figure 2. The lag τ and lag difference δ are equal.

From (7), we have that b_1 is independent from a_1 and b_1 is independent from b_2. Thus, we have that the information b_n injected to X_n and Y_n from L is an i.i.d. sequence. A similar argument shows that the information c_1 passed from $Y_{n+\tau}$ to $Y_{n+\tau+1}$ is independent from b_2. We still cannot decide if X drives Y or L drives both; however, in the latter situation, we may say that L emits observational noise for X, and it does not influence its evolution (the value of a_i). Alternatively, we may consider b_i as the "part" of X, which is injected to Y. Let us note that L itself is not necessarily an i.i.d. sequence but, from the point of view of its impact on X and Y, it is indifferent.

One may appeal to the Occam's razor principle (if other background knowledge does not dictate otherwise) that L itself is an i.i.d. process. If b is part of X or external noise that cannot be decided without further knowledge, we may refer again to the Occam's razor principle and assume that there is no a third system, a common driver but X injects an i.i.d. sequence to Y.

2.3. Geodesic Spaces

Now, we investigate the case when the subsystems of M are located in a geodesic metric space with unique geodesics between any pair of points. We assume that the information transfer speed is uniform, constant in the space regardless of the location of the source and target. Under that assumption, we can speak interchangeably about distance in space and time.

2.4. Strict Reversed Triangular Inequality

If $\delta = \min_L \delta_{L,X,Y}$ and

$$\delta > \tau \quad (8)$$

then we have

$$\tau_{L,Y} > \tau_{L,X} + \tau_{X,Y} \quad (9)$$

the reversed, strict triangular inequality, and there is information share between X_n and $Y_{n+\tau}$, then no L can be a common driver of X_n and $Y_{n+\tau}$ (cf. Figure 3), so a direct driving should take place from X to Y.

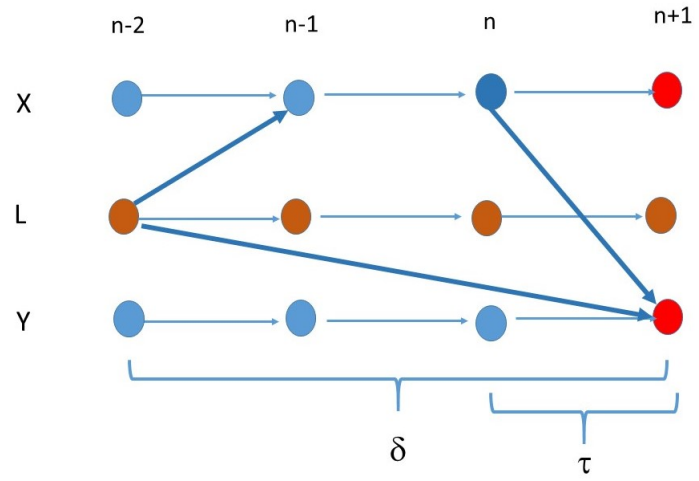

Figure 3. The causation has smaller time lag τ compared with the difference δ from the common driver.

2.5. Strict Triangular Inequality

On the other hand, if for an L

$$\delta_{L\,X,Y} < \tau_{X,Y} \tag{10}$$

then we have

$$\tau_{L,Y} < \tau_{L,X} + \tau_{X,Y} \tag{11}$$

and X_n and $Y_{n+\tau}$ have positive conditional mutual information conditioned on the past, then only L, the common driver can produce it, not causation (see Figure 4).

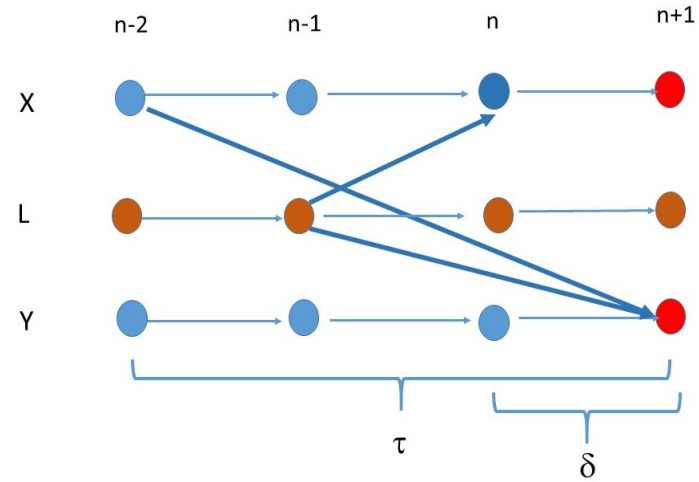

Figure 4. The causation has larger time lag τ compared with the difference δ from the common driver.

2.6. The Equality, the Confounder Case

Finally, if

$$\tau_{X,Y} = \delta_{L,X,Y},$$
$$\tau_{L,Y} = \tau_{L,X} + \tau_{X,Y} \tag{12}$$

we have a confounder.

If the metric space has a unique geodesic from L to Y, then X should be on that geodesic of L and Y, and this means that the information from L either enters X along the path to Y or avoids it in in a tricky way by an infinitesimal detour as Figure 5 depicts.

Figure 5. The causation time lag τ equal to the lag difference δ.

In the former case, we have no confounder but the causal chain $L \to X \to Y$. This is a situation that, again, cannot be resolved without additional information about the actual systems under scrutiny. Economists used to call such L an instrumental variable.

Now, let us recall that the inequalities (8) and (10) read as

$$\tau_{L,Y} > \tau_{L,X} + \tau_{X,Y}, \tag{13}$$

$$\tau_{L,Y} < \tau_{L,X} + \tau_{X,Y}. \tag{14}$$

The latter one is the strict triangular inequality and the former one is its converse (both with strict inequality). Here, we arrive at the interpretation of causation in M. If it is a system in an abstract space without metric properties, there is no point to speak about distances in it, and there is no link between information transfer time (lag in short) and distances.

On the other hand, if

- the system M is located in a geodesic metric space,
- the geodesics are unique,
- the information propagates along the geodesics, and
- the information transfer has a constant speed,

then , distances are proportional to the delay with the same constant factor for all members. Triangular inequality is inherited from distances to lags. In the case of a metric space, like the Euclidean, hyperbolic and spherical with unique geodesic (except if X and Y are the oppositely positioned on the sphere) the triangular inequality holds, and thus (13) is impossible, and L cannot be a common driver that mimics driving or acts parallel to a driving between X and Y. Let us note that the triangular inequality holds for space-like vectors in the Minkovsky space, while the converse holds for time-like positions. Finally, the case of strict equality needs further investigation.

In case of different transfer speed, the picture is more complex, and the above geometric consideration is applicable in particular settings only. In the human brain, the information transfer has different speed depending on the transfer mode: via sequences of neural cells, long axon bundles or volume of surface currents. The transfer speed depends on the number of intermediate relay nodes of the network as well. Consequently, the case of causality analysis of brain regions needs detailed information on the connection type and speed between them. It is likely that many other topical areas, like climate and geophysics, specific knowledge of the metric properties and transfer speed may contribute to the success of causal discovery. In other areas, there is no information about the temporal arrangement of the unobserved factors, and consequently revealing the perfect description of the causal structure seems impossible.

2.7. Conditions and Mixing

Let us recall here that all the methods that are based on Pearl's DAG analysis use d-separation (or causal Markovness) based on a conditional independence test (CIT) in which parents are the conditioning variables. As such, they need access to the parents, which is impossible if those are not observed, and the computation cost can be prohibitive for large networks. Let us see that the d-separation uses the parents as cut set in the DAG. In Section 2, we used the full past of both observed processes. In practice, it is

impossible to put the whole past in the condition; therefore, we should work with a shorter history. Let us consider, as an example, the case when $0 \leq k < \tau_{xy}$, which means that there is no information transfer from x_n to y_{n+k} and investigate $I\left(x_n\,; y_{n+k}|X_0^{n-1}, Y_0^{n+k-1}\right)$. If $I\left(x_n\,; y_{n+k}|X_0^{n-1}, Y_0^{n+k-1}\right) = 0$, i.e., there is no hidden common driver. One can show that

$$I\left(x_n\,; y_{n+k}|X_0^{n-1}, Y_0^{n+k-1}\right) \leq I(x_n\,; y_{n+k}|J_d) \to 0$$

as $d \to \infty$, where $J_d = \left(X_{n-d}^{n-1}, Y_{n\,-d}^{n+k-1}\right)$. For the proof, see Appendix A.

With this argument, we have that the convergence to a constant or to zero of the conditional mutual information determines if there is a driving between X and Y and if there is a common driver (as indicated in Relation 1). Under Assumption 5, it is evident that if there is a hidden common driver, the information is passed along a fixed length path from the common cause to X and Y, and its effect on dependency is not diminishing. If there is no common driver, the exchanged information should traverse longer and longer paths, and the Conditional Mutual Information (CMI) should go to zero as d goes to infinity.

The conditional independence test (and proper estimate of CMI) has recently been the focus of research motivated by applications in machine learning and artificial intelligence. This is known to be a challenging task (cf. [6,12,17–19]).

3. Related Works and Discussion

There are numerous extensions and refinements of the original PC algorithm that Pearl developed [20]. This applies to the study of causal discovery of dynamic systems based on observed time series. We mention some prominent works [4,6,12,21,22]and their bibliography for further reading (see also the extended surveys [5,23]). The recent works [12,14] (see also [23]) have a very similar approach to the present one. In particular, we also use the structural modeling framework; however, we limit our focus to the discovery of a causal relation between a pair of systems. The method can be extended to the study of many time series by considering vector valued observations and/or many pairwise investigations.

The capabilities and limitations of the causal discovery algorithms were investigated in detail in seminal works [15,16,20,24] and recently in [14,21]. The recent generalizations are complete. They extend the labeling of edge ends of classical DAGs, while completeness does not mean that all relations are well specified. Completeness means that all the possible MAGs (Markov Equivalent Acyclic Graphs) can be created.

In this paper, we used an essential assumption and two unavoidable approximations. First, we assumed that the continuous time process can be inferred using a discrete time and limited resolution time series observation. Next, we assumed that the discrete time process can be well approximated with an order-p SVAR model. Finally, if the processes contain continuous variables, the condition is not restricted to a single state value but to a set of them, and, as a consequence, it is not perfectly blocking the information flow between the marginal variables. This deficiency might be eliminated by the local permutation method proposed by Runge in [19].

Author Contributions: Conceptualization, Á.Z. and A.T.; Formal analysis, A.T.; Methodology, Á.Z. and A.T.; Writing—original draft, A.T.; Writing—review and editing Z.B., M.S. and Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: Z.B., Z.S. and A.T. was partially supported by an award from the National Brain Research Program of Hungary (NAP-B, KTIA NAP 12-2-201), Z.S. Hungarian National Research, Development and Innovation Fund, NKFIH under grant number K 135837.

Data Availability Statement: Not applicable.

Acknowledgments: We thank to Roberta Rehus for editing the paper.

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

Appendix A

We show that if

$$I\left(x_n\,;y_{n+k}|X_0^{n-1},Y_0^{n+k-1}\right)=0 \tag{A1}$$

it can be confirmed by collecting evidence that

$$I\left(x_n\,;y_{n+k}|X_{n-d}^{n-1},Y_{n-d}^{n+k-1}\right)\to 0$$

as d tends to infinity. Let us assume that (A1) holds. Since $k<\tau_{x,y}$ and there is no common cause of any information, what is shared by x_n and y_{n+k} should come from their past.

Let us introduce the short notation $J_d=\left(X_{n-d}^{n-1},Y_{n\,-d}^{n-1}\right)$ and estimate $I\left(x_n\,;y_{n+k}|X_0^{n-1},Y_0^{n+k-1}\right)$ from above using the monotonicity of the conditional mutual information.

$$\begin{aligned} I\left(x_n\,;y_{n+k}|X_0^{n-1},Y_0^{n+k-1}\right) &\leq I\left(x_n\,;y_{n+k}|X_{n-d}^{n-1},Y_{n-d}^{n+k-1}\right)\\ &\leq I\left(X_n^{n+k};Y_n^{n+k}|J_d\right). \end{aligned}$$

the first step, we use the monotonicity, then the two assumption and monotonicity again. Next, we use that shared information comes from the past then monotonicity again in the second steps:

$$\begin{aligned} I\left(X_n^{n+k};Y_n^{n+k}|J_d\right) &\leq I\left(M_n^{n+k};M_0^{n-1-d}|J_d\right)\\ &\leq I\left(M_n^{n+k};M_0^{n-1-d}\right). \end{aligned}$$

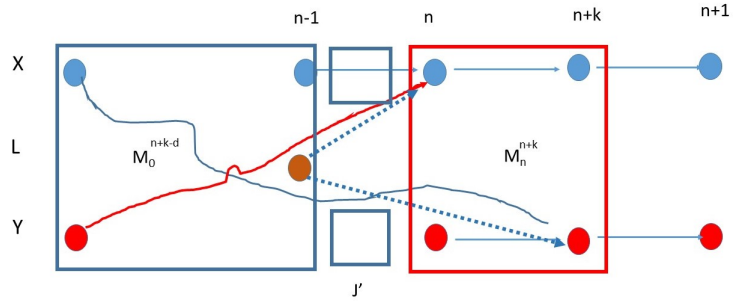

Figure A1. The figure shows why the conditions do not block the common driver.

Now, we use the fact that m is a first order Markov chain, then that it is time homogeneous and finally that, from exactness, it follows that it is α-mixing.

$$\begin{aligned} & I\left(x_n\,;y_{n+k}|X_0^{n-1},Y_0^{n+k-1}\right)\\ \leq\ & I\left(M_n^{n+k};M_0^{n-1-d}\right)\\ =\ & I(M_{d+1}\,;M_0)\leq\alpha(d+2)\to 0 \end{aligned}$$

as $d\to\infty$ due to the strong mixing property, following from the exactness assumption (5).

References

1. Wiener, N. The theory of prediction. In *Modern Mathematics for Engineers*; Beckenbach, E., Ed.; McGraw-Hill: New York, NY, USA, 1956.
2. Granger, C. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
3. Maziarz, M. A review of the Granger-causality fallacy. *J. Philos. Econ. Reflections Econ. Soc. Issues* **2015**, *8*, 86–105.

4. Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Coumou, D.; Deyle, E.; Glymour, C.; Kretschmer, M.; Mahecha, M.D.; Muñoz-Marí, J.; et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **2019**, *10*, 2553. [CrossRef] [PubMed]
5. Guo, R.; Cheng, L.; Li, J.; Hahn, P.R.; Liu, H. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv. CSUR* **2020**, *53*, 1–37. [CrossRef]
6. Guyon, I.; Janzing, D.; Schölkopf, B. Causality: Objectives and assessment. In *Causality: Objectives and Assessment*; PMLR: Maastricht, The Netherlands, 2010; pp. 1–42.
7. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.H.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500. [CrossRef] [PubMed]
8. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick, 1980*; Springer: Berlin/Heidelberg, Germany, 1981; pp. 366–381.
9. Stark, J. Delay embeddings for forced systems. I. Deterministic forcing. *J. Nonlinear Sci.* **1999**, *9*, 255–332. [CrossRef]
10. Stark, J.; Broomhead, D.S.; Davies, M.E.; Huke, J. Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.* **2003**, *13*, 519–577. [CrossRef]
11. Benko, Z.; Zlatniczki, A.; Fabó, D.; Sólyom, A.; Erõoss, L.; Telcs, A.; Somogyvári, Z. Complete inference of causal relations in dynamical systems. *arXiv* **2018**, arXiv:1808.10806.
12. Malinsky, D.; Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In Proceedings of the 2018 ACM SIGKDD Workshop on Causal Discovery, London, UK, 20 August 2018; PMLR: Maastricht, The Netherlands, 2018; pp. 23–47.
13. Lasota, A.; Mackey, M.C. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 97.
14. Mastakouri, A.A.; Schölkopf, B.; Janzing, D. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. *arXiv* **2020**, arXiv:2005.08543.
15. Zhang, J. Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.* **2008**, *9*, 1437–1474.
16. 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. [CrossRef]
17. Li, C.; Fan, X. On nonparametric conditional independence tests for continuous variables. *Wiley Interdiscip. Comput. Stat.* **2020**, *12*, e1489. [CrossRef]
18. Lundborg, A.R.; Shah, R.D.; Peters, J. Conditional Independence Testing in Hilbert Spaces with Applications to Functional Data Analysis. *arXiv* **2021**, arXiv:2101.07108.
19. Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Beijing, China, 18–20 August 2018; PMLR: Maastricht, The Netherlands, 2018; pp. 938–947.
20. Pearl, J. *Causality*; Cambridge University Press: Causality, UK, 2009.
21. Lin, H.; Zhang, J. January. On Learning Causal Structures from Non-Experimental Data without Any Faithfulness Assumption. In *Algorithmic Learning Theory*; PMLR: Maastricht, The Netherlands, 2020; pp. 554–582.
22. Sun, J.; Taylor, D.; Bollt, E.M. Causal network inference by optimal causation entropy. *SIAM J. Appl. Dyn.* **2015**, *14*, 73–106. [CrossRef]
23. Vowels, M.J.; Camgoz, N.C.; Bowden, R. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv* **2021**, arXiv:2103.02582.
24. Spirtes, P.; Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* **1991**, *9*, 62–72. [CrossRef]