mathematics

MDPI

Article

RCDi: Robust Causal Direction Inference Using INUS-Inspired Asymmetry with the Solomonoff Prior

Ling Zhao, Zhe Chen, Qinyao Luo, Silu He and Haifeng Li *

School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; zhaoling@csu.edu.cn (L.Z.); 235011047@csu.edu.cn (Z.C.); qinyaoluo@csu.edu.cn (Q.L.); hesilu@csu.edu.cn (S.H.)
* Correspondence: lihaifeng@csu.edu.cn

Abstract: Investigating causal interactions between entities is a crucial task across various scientific domains. The traditional causal discovery methods often assume a predetermined causal direction, which is problematic when prior knowledge is insufficient. Identifying causal directions from observational data remains a key challenge. Causal discovery typically relies on two priors: the uniform prior and the Solomonoff prior. The Solomonoff prior theoretically outperforms the uniform prior in determining causal directions in bivariate scenarios by using the causal independence mechanism assumption. However, this approach has two main issues: it assumes that no unobserved variables affect the outcome, leading to method failure if violated, and it relies on the uncomputable Kolmogorov complexity (KC). In addition, we employ Kolmogorov's structure function to analyze the use of the minimum description length (MDL) as an approximation for KC, which shows that the function class used for computing the MDL introduces prior biases, increasing the risk of misclassification. Inspired by the insufficient but necessary part of an unnecessary but sufficient condition (INUS condition), we propose an asymmetry where the expected complexity change in the cause, due to changes in the effect, is greater than the reverse. This criterion supplements the causal independence mechanism when its restrictive conditions are not met under the Solomonoff prior. To mitigate prior bias and reduce misclassification risk, we introduce a multilayer perceptron based on the universal approximation theorem as the backbone network, enhancing method stability. Our approach demonstrates a competitive performance against the SOTA methods on the TCEP real dataset. Additionally, the results on synthetic datasets show that our method maintains stability across various data generation mechanisms and noise distributions. This work advances causal direction determination research by addressing the limitations of the existing methods and offering a more robust and stable approach.

Keywords: causal discovery; minimum description length; regression

MSC: 62D20

check for updates

Academic Editors: Bumba Mukherjee and Marjan Mernik

Received: 14 January 2025
Revised: 27 January 2025
Accepted: 3 February 2025
Published: 6 February 2025

Citation: Zhao, L.; Chen, Z.; Luo, Q.; He, S.; Li, H. RCDi: Robust Causal Direction Inference Using INUS-Inspired Asymmetry with the Solomonoff Prior. *Mathematics* **2025**, *13*, 544. https://doi.org/10.3390/math13030544

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The era of big data has facilitated great strides in scientific research. However, traditional research methods based on correlations, such as machine learning techniques, often lead to knowledge containing spurious correlations, which can lead to erroneous perceptions of the world [1]. Effectively and accurately obtaining causal relationships is more practically significant than merely identifying and detecting correlations. For example, there is a relationship between weather temperature and ice cream sales [2]. A prediction model for ice cream sales based on correlations cannot further guide ice cream shop owners

in devising more appropriate sales strategies. Similarly, the same temperature can result in completely different sales figures at the beach compared with urban office areas. Therefore, we are more concerned with how temperature causally affects ice cream sales after other potential confounding variables are removed. Randomized controlled trials (RCTs) are widely regarded as the gold standard for studying causal relationships [3]; however, it is worth noting that RCTs implicitly assume the causal direction between the two variables. The design of RCTs is based on researchers having already determined which variable is the potential "cause" (intervention variable) and which is the "effect" (outcome variable). Unfortunately, such knowledge about the direction between two variables is difficult to obtain in many research fields, such as the widely discussed climate warming and carbon dioxide concentration [4], and the causal directions in relationships such as economic growth and technological innovation [5]. Therefore, inferring causal directions from observational data is a significant research problem in causal science.

Generally, we consider causal relationships to be a subset of correlation relationships. The uniqueness of causal relationships lies in their logical asymmetry, meaning that we find it more reasonable to intervene in the cause to induce changes in the effect rather than to intervene in the effect to cause changes in the cause [6]. This asymmetry implies that causal relationships are directional, so we typically use directed acyclic graphs (DAGs) to describe the causal relationship structure between variables [7]. We need to utilize certain fundamental priors to constrain and filter which correlations can be considered causal relationships or, in other words, to determine under what circumstances we can regard these correlations as containing causal relationships. Identifying causal relationships from observational data is referred to as causal discovery. In the field of causal discovery, there are two fundamental priors: the uniform prior and the Solomonoff prior [8]. Although the former may not have been directly mentioned, one of its implications, the minimality of factorization, is widely used in the field of causal structure discovery. In short, in directed acyclic graphs (DAGs) that describe the conditional independence of variables, the true causal structure should be the DAG with the fewest edges. However, the task of causal direction identification differs from the general causal structure discovery tasks; we only have data for two variables, and therefore, the commonly used conditional independence tests in causal structure discovery cannot be implemented [9]. Such a predicament drives us to further seek assumptions about causal generative mechanisms, for example, the causal independence mechanism (ICM) assumption based on the Solomonoff prior, which is based on the idea that the simplest explanation or model for a given observation should be preferred. This assumption posits that the mechanism generating the cause variable and the cause–effect mechanism are mutually independent [10]. On the basis of these assumptions, it becomes possible to identify the asymmetrical relationships between causes and effects as statistical asymmetries, from the data, thereby allowing us to infer causal directions using observational data [11].

The existing theories indicate that the Solomonoff prior is more suitable for cases involving observational data with only two variables [12]. This point can be validated by the fact that many causal direction determination methods based on the ICM assumption (which follows the Solomonoff prior), such as SLOPE [13], achieve better performance on commonly used TCEP datasets compared to methods based on structural equation models, such as ANM [14]. However, the widespread validity of the causal independence mechanism assumption based on the Solomonoff prior remains to be verified [15]. To enhance the application of the Solomonoff prior in causal direction inference, we explore an assumption that, while still based on the Solomonoff prior, holds outside the applicability range of the causal independence mechanism, thereby establishing a new principle for causal direction inference. Our analysis identifies one of the primary limitations of the

causal independence mechanism: it requires that all parent nodes of the outcome variable in the causal relationship be fully included within the observational dataset [8]. Using the task of inferring the causal direction between two variables as an example, if a causal relationship X→Y exists, the principle of inferring the causal direction based on the causal independence mechanism is valid under the condition that Y does not have any unobserved parent nodes other than X. To overcome this limitation, we introduce the insufficient but necessary part of an unnecessary but sufficient condition (INUS condition) [16], a concept in causation theory, as our fundamental assumption. If the causal mechanism conforms to the form defined by the INUS condition, we propose that there exists an asymmetry in the expected complexity changes between the cause and the effect. On this basis, we design our causal direction inference principle.

When instantiating our principle, since we adhere to the Solomonoff prior, the complexity within the principle corresponds to the uncomputable algorithmic complexity. Currently, this is generally approximated using the two-part minimum description length (two-part MDL). The existing methods typically construct a set of basis functions and calculate the final two-part MDL score by statistically balancing the types and quantities of basis functions used by the model along with the fitting accuracy. Through Kolmogorov's structure function [17], we theoretically find that constructing the function class used to compute the two-part MDL may introduce prior biases, ultimately leading to the risk of misclassification caused by these prior biases when the causal direction inference principle is applied. To address this issue, our method employs a multilayer perceptron (MLP) based on the universal approximation theorem as the backbone network of our model. By integrating indicators that evaluate the model's complexity based on the weight values with the fitting accuracy into the loss function [18] and balancing these two components through gradient descent learning, our approach mitigates the risk of misclassification, and enhances the method's stability. This is achieved by eliminating the need to manually construct a set of basis functions, thereby reducing the potential for prior-induced errors, and improving the overall robustness.

Our method achieves competitive accuracy compared with the existing state-of-the-art (SOTA) methods on the widely used real bivariate causal discovery dataset, TCEP. Furthermore, unlike some methods that typically require univariate data as input, another advantage of our approach is its ability to handle all causal pairs within the dataset comprehensively. Additionally, experiments on commonly used synthetic datasets demonstrate that our method maintains excellent stability across a variety of data generation mechanisms with different functional forms.

Our contributions can be summarized in the following two points:

- A new causal direction inference principle is proposed on the basis of the INUS condition. The limitations of the causal independence mechanism assumption is overcome in causal direction inference tasks (such as the presence of unobserved parent nodes for the outcome variable). By introducing an asymmetry in the complexity changes accompanying the causal directions, we enhance the applicability of the Solomonoff prior.
- We develop an MLP-based MDL scoring mechanism for computing our causal direction inference principle. By adopting a function class-free approach, we avoid the prior bias that may arise from the manual construction of basis function sets in the traditional methods, thereby enhancing the stability and accuracy of the causal direction inference. Furthermore, we theoretically demonstrate the necessity of this approach.

The remainder of this paper is organized as follows. We first introduce related work and some preliminary knowledge. The methods section primarily consists of two parts: first, we outline the new proposed causal inference principle, where we also formally define

the causal direction determination problem; second, we detail the specific implementation of the algorithm. Finally, the experimental section includes both real and synthetic data.

2. Related Works

In the task of determining the causal direction between two variables, many classical methods have been widely applied across various research fields [2,19–21], such as Granger causality [22] and the convergent cross mapping (CCM) [23] methods. However, these methods are designed for time-series data [24], and their effectiveness relies on high-quality time series observational data. Our method addresses the problem of causal direction determination for non-time-series data, as in many cases—such as in the field of earth sciences—long time-series data are difficult to obtain, or are prohibitively costly to acquire [25].

The existing methods for causal direction determination in non-time-series data can be broadly categorized into two types based on the specific location of their assumption for obtaining statistical asymmetries:

1. Methods based on structural equation models: These approaches utilize structural equations to model the causal relationships between variables, leveraging the inherent structure to infer causality.
2. Methods based on causal independence mechanisms: These approaches rely on the assumption that the causal mechanisms of the cause and effect are independent, using this independence to determine the direction of causality.

The function model-based approach posits that causal relationships between two variables can be represented using a function model. For example, if $X \to Y$, there exists a fitted function model $Y = g(f(X), n)$, where n is a noise variable independent of X. Conversely, if we fit a model in the opposite direction, $X = g(f(Y), n')$, the noise variable n' will not be independent of Y. In practical implementation, additional assumptions regarding the specific forms of f and g and the distribution of n are often needed. For example, the LiNGAM method [26] assumes that the causal relationship follows a linear function model, where f is linear, and the noise variable n follows a Gaussian distribution. Furthermore, a series of approaches have been developed to relax these assumptions to broaden their applicability. For example, ANM [14] does not assume linearity in the causal relationship but assumes that the noise is additive. On the other hand, PNL [27] permits the additive noise to combine with the causative variable after a nonlinear transformation, which is then passed through another nonlinear function to generate the outcome variable. These methods share the common principle of determining the causal direction by testing the independence of the noise and the independent variable, and their effectiveness is contingent upon the data-generating mechanism satisfying the specific assumptions of the function model [9]. Recent methods, such as CEMM [28], leverage a learnable module to integrate these different model-based assumptions, thereby ensuring the selection of the most appropriate function model under varying data-generating mechanisms.

Another category of methods, which is based on the causal independence mechanism hypothesis, does not require additional assumptions about the specific functional form of the causal relationship. These methods adhere to the Solomonoff prior, which posits that the mechanism governing the causal relationship is independent of the mechanism generating the cause, whereas the mechanism generating the effect is not independent of the causal mechanism, and involves algorithmic mutual information. Consequently, their fundamental principle for determining the causal direction remains consistent, expressed as $I(X : Y|X) = 0$, or equivalently $K(Y|X) \leq K(Y) + K(X|Y)$. However, the specific approaches used to instantiate the uncomputable K differ among these methods. For example, Origo [29] uses binary string data as input and approximates K by measuring the diffi-

culty of lossless compression. On the other hand, CURE [30] employs a semisupervised inverse regression technique to assess whether the marginal distributions contain information about the joint distribution, thereby directly representing the presence of algorithmic mutual information. SLOPE [13], which follows the two-part MDL principle, uses the complexity of the regression model and regression loss, achieving state-of-the-art (SOTA) performance on widely used real-world datasets, such as TCEP [31]. In fact, while methods based on the causal independence mechanism no longer rely on restrictions regarding the functional form of the regression model between causal variables, most of these methods still approximate K through regression tasks.

3. Preliminaries

In this section, we briefly introduce the algorithmic information theory concepts involved in NMDL4CO, and the INUS condition assumptions used to construct our causal direction inference principles.

3.1. Kolmogorov Complexity

Kolmogorov complexity, also known as algorithmic complexity, is used to quantify the complexity of an object [17]. It can be defined as the length of the simplest description of an object. In other words, for a given string s, its Kolmogorov complexity $K(s)$ is the length of the shortest program capable of generating s (on a fixed universal computational model, such as a Turing machine). More formally, it can be defined as:

$$K(s) = min_{p:U(p)=s}|p| \tag{1}$$

where s is the string to be described. p is the program input to a universal Turing machine U such that $U(p)$ outputs the string s. $|p|$ represents the length of the program (in bits).

This paper also involves an extension of Kolmogorov complexity, known as conditional Kolmogorov complexity, which considers the length of the shortest program required to describe an object given some additional information. Its definition is similar to that of Kolmogorov complexity:

$$K(s \mid t) = min_{p:U(p,t)=s} |p| \tag{2}$$

t is the conditional string. p is a program input to a universal Turing machine U such that, given the condition t, the program p outputs the string s. For example, suppose that we have two strings s and t:

$$\text{s} = \text{“abababababababab”t} = \text{“ab”}$$

In this case, the conditional Kolmogorov complexity of s given t, denoted as $K(s \mid t)$, would be relatively low, because s can be described using a simple repetitive pattern that is based on t: “repeat t seven times.”

3.2. Minimum Description Length

For any string s, its Kolmogorov complexity $K(s)$ is uncomputable because there is no universal method to find the shortest program that generates s. The MDL principle aims to minimize the description length of the data by selecting the optimal model. This aligns with the goal of Kolmogorov complexity: finding the most concise description. Thus, MDL can be used as an approximation of Kolmogorov complexity. Specifically, MDL can be defined in the following two-part form:

$$\text{MDL}_{\text{two-part}} = L(\text{model}) + L(\text{data} \mid \text{model}) \tag{3}$$

- Model description length $L(\text{model})$: Select a model and calculate the encoding length required to describe it. This corresponds to selecting an appropriate program with Kolmogorov complexity.
- Data description length $L(\text{data} \mid \text{model})$: Given the model, calculate the length required to describe the data. This part is analogous to running the program to generate the data.

In practical computations, to ensure computability, the MDL principle relaxes the requirement for Turing completeness. In simple terms, MDL does not aim to search for the optimal model across all possible model sets but instead approximates by searching for the best model within a predefined model search space.

3.3. INUS Condition (Insufficient but Necessary Part of an Unnecessary but Sufficient Condition)

For an event A, if events X and Y exist such that $(A \cap X) \cup Y$ is a necessary and sufficient condition for the occurrence of event E, and A alone is not sufficient to cause E but merely constitutes an unnecessary, yet sufficient, part of $A \cap X$, where $A \cap X$ itself is an insufficient but necessary part of $(A \cap X) \cup Y$, then A is called an INUS condition for E [16].

The primary significance of the INUS condition lies in partially completing the full meaning of causal statements. It emphasizes the factors involved in constituting causal relationships rather than focusing on the independent mechanisms that highlight the independent pathways of causal connections.

An example of an INUS condition can be illustrated with the scenario of a house fire. Suppose a cigarette is carelessly discarded, and it causes a fire. However, the cigarette alone is not sufficient to start the fire; it requires the presence of a flammable material, such as dry curtains, to ignite. The combination of the cigarette and the flammable material is sufficient to cause the fire, but neither one is enough by itself. In this case, the cigarette is an insufficient condition, but it is necessary as part of the combination. So, we can say that a carelessly discarded, still-lit cigarette is an INUS condition for a fire.

4. Methodology

4.1. Problem Description

This paper focuses on the problem of determining the causal graph between two observed variables, X and Y, specifically whether the causal direction is $X \to Y$ or $X \leftarrow Y$. In fact, we assume that X and Y are statistically dependent, and that this statistical dependence arises from a causal effect in the form of either $X \to Y$ or $X \leftarrow Y$. This task can be regarded as a special case of the causal discovery problem, which we refer to as causal orientation. Table 1 summarizes the main mathematical symbols and their meanings.

Table 1. List of symbols.

Symbol	Meaning	Symbol	Meaning
$K(X)$	X's Kolmogorov complexity	S_c	Minimal sufficient set of C
$L(X)$	X's MDL	C_1	One of the causal variables of E
$M_{X\|Y}$	A model that encodes X, given Y	c_1	A sample value of C_1
$D_{X\|Y}$	Data X encoded, given Y	E	Outcome variable of interest
$P(X \mid Y)$	Conditional distribution	e	A sample value of E
$P(X)$	X's distribution	I	Defined event
P(I)	Probability of I occurring		
$\mathbb{E}(X)$	Expected value of X		

4.2. The Causal Orientation Principle Inspired by the INUS Condition

Motivation

Solomonoff's prior is an elegant implementation of Occam's Razor, which posits that simple generative mechanisms are more likely to occur than complex, random ones. This prior can serve as an insight into causal mechanisms, suggesting that in a system, if we describe the causal interactions between elements using a causal graph structure, a graph with lower complexity is more likely to represent the true causal relationships. We typically use Kolmogorov complexity as a measure of the complexity of such mechanisms. Existing theories have proven that this prior is more suitable for the causal orientation task than another commonly used prior in causal discovery, the uniform prior.

Most methods following Solomonoff's prior adhere to the independent causal mechanism (ICM) assumption. This assumption posits that $K(s_1, \ldots, s_n) \ll \sum_{j=1}^{n} K(s_j)$, meaning that in a system with causal relationships, using causal mechanisms for a joint description results in a lower complexity than the sum of the complexities of describing each element individually. Specifically, $K(P(X_1, \ldots, X_n)) \stackrel{+}{=} \sum_{i=1}^{n} K(P(X_i \mid \text{Parents}(X_i)))$. Implicit in this process is the assumption that the parent set of X_i is known. In the case of a causal orientation, where only two observed variables are present, we have $K(P(X,Y)) \stackrel{+}{=} K(P(X)) + K(P(Y \mid X))$ or $K(P(X,Y)) \stackrel{+}{=} K(P(Y)) + K(P(X \mid Y))$. This implies that regardless of whether the true causal graph is $X \to Y$ or $X \leftarrow Y$, the IC assumption inherently presumes that the parent set of the effect variable includes only the other variable, i.e., "no other unobserved causes exist." However, this assumption is often difficult to verify in real-world scenarios. For example, the average life expectancy of residents in a region clearly depends on numerous causal variables, many of which may be unobservable. It should be noted that the other unobserved variables we mentioned above do not imply that they are confounding variables. From a causal graph perspective, the unobserved variable we refer to is not a parent node of the observed cause variables, whereas a confounding variable is a parent node of both the cause and outcome variables. To overcome this limitation, we propose a new principle for causal orientation inspired by the INUS condition, The main difference between our method and the ICM-based methods is shown in Figure 1.

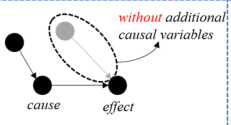
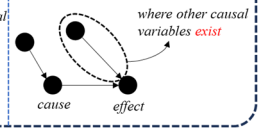
	ICM-Based	**Our Approach**
Basic Prior	Solomonoff Prior	
Assumption	Causal IC Condition	INUS Condition
Formula (Inferring Score)	$K(P(X)) + K(P(Y \mid X))$	$\frac{\mathbb{E}\left(K(P(C_1\|e))\right)}{K(P(C_1))}$
Property		

Figure 1. The difference between the causal orientation principles of our method and the ICM-based methods.

In this work, we assume that if variable X is an INUS condition for Y, then X can be regarded as a causal variable for Y. Since the INUS condition does not require the causes corresponding to an effect to be unique, our principle does not assume that the variable identified as the cause is the sole cause of the effect. Instead, we allow for the possibility of

other unobserved causes. This stands in contrast to the ICM framework, which implicitly assumes the absence of other unobserved causes. This distinction represents the most significant difference between our approach and the other ICM-based methods.

4.3. Theoretical Derivation and Formula

We use Figure 2 to specifically illustrate the application of the INUS condition. Dashed arrows represent necessary condition relationships, while solid arrows represent sufficient condition relationships. Here, C is the variable considered the cause in the INUS condition, E is the variable regarded as the effect, and S_c represents the minimal sufficient set of variables for causing a change in E, with C being a necessary condition within this set. The solid arrows in the diagram indicate sufficient condition relationships, whereas the dashed arrows represent necessary condition relationships. Among the variables C, S_c, and E, there exist events where a change in one variable is accompanied by a change in another. Specifically, we focus on the four types of events depicted in the diagram, and the probability relationships of these events form the foundation for deriving our causal orientation principle.

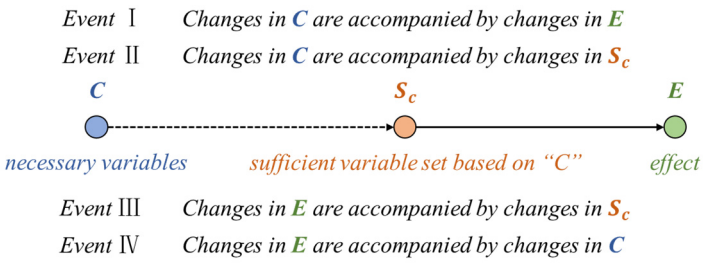


Figure 2. Introduction to the INUS Condition.

Both C and E are necessary conditions for S_c. Since S_c is defined as the minimal sufficient set for C under the INUS condition, C is the maximal necessary subset of S_c. Any other necessary condition set, including E, has a sufficiency less than or equal to that of C. Therefore, we can observe that the probability of Event I occurring is greater than that of Event III. On the other hand, if we analyze the specific situations in which each event occurs, the occurrence of Event I necessarily implies the occurrence of Event II. This is because, according to the definition of the INUS condition, a change in S_c is a sufficient condition for a change in E. Conversely, the occurrence of Event II does not necessarily imply Event I. A change in E might be caused by another sufficient set outside of S_c, with C coincidentally changing at the same time. Similarly, $\neg \mathrm{III} \bigvee \mathrm{IV}$ reflects the same reasoning. Now, we can establish the following probability relationships: $\mathrm{P(II)} > \mathrm{P(I)} > \mathrm{P(III)}$ and $\mathrm{P(IV)} - \mathrm{P(III)} = \mathrm{P(II)} - \mathrm{P(I)}$. Thus, we have:

$$P(\mathrm{II}) \geq P(\mathrm{IV}) \tag{4}$$

Since we aim to derive a general causal relationship (type/general causality) between classes of events that is independent of the specific event contexts, we focus on how the distribution of one variable changes when the distribution of another variable changes, rather than the specific values of the variables. This implies that when using Kolmogorov complexity to construct the inference principle, the objects we measure are the distributions P(C) and P(E), specifically their conditional complexities. By incorporating Event II and substituting specific events into Equation (4), we have:

$$\frac{\mathbb{E}[K(P(C_1|e))]}{K(P(C_1))} > \frac{\mathbb{E}[K(P(E|c_1))]}{K(P(E))} \tag{5}$$

For simplicity, if we assume that the initial complexities of the cause and effect are comparable, i.e., the differences in algorithmic complexity between the marginal distributions of the two variables are not significant enough to fundamentally affect the causal direction inference, then we can infer the causal direction solely by comparing the numerators of the equation. The most intuitive way to model the numerator is to compute the complexity of the distribution of the effect (or cause) for each fixed value of the cause (or effect) based on the observed data. However, this approach is clearly impractical, because for the continuous observational data, the samples corresponding to a fixed value of a variable are often extremely few or even unique. To address this, we approximate the expected algorithmic complexity of the effect (or cause) distributions across all values of the cause (or effect) using conditional algorithmic complexity. Given the uncomputability of algorithmic complexity, we substitute it with the two-part MDL score of a regression model.

Our causal direction inference rule can be formalized as the following equation:

$$MDLScore_{X\to Y} = \alpha\left(L\left(M_{Y|X}\right)\right) + \beta\left(L\left(D_{Y|X,x}\right)\right) \tag{6}$$

$$MDLScore_{Y\to X} = \alpha\left(L\left(M_{X|Y}\right)\right) + \beta\left(L\left(D_{X|Y,y}\right)\right) \tag{7}$$

If:

$$MDLScore_{X\to Y} - MDLScore_{Y\to X} > 0 \tag{8}$$

We determine X as the cause of Y if the above condition holds and, conversely, Y as the cause of X otherwise. Here, $L(M)$ and $L(D)$ represent the complexity of encoding the model and the complexity of encoding the data after applying the model, respectively. The coefficients α and β adjust the relative importance of these two components in the overall *MDL score*.

4.4. Algorithm Implementation for Our Causal Orientation Principle

We model the conditional distribution of Y given X using a regression approach. Specifically, we employ a multilayer perceptron (MLP) as our regression model, which does not require the construction of a predefined function class. The neural MDL (NMDL) score is used to measure the complexity of encoding the model, whereas regression loss is employed to quantify the complexity of encoding the data after applying the model. The combination of model simplicity and regression error forms the loss function, which is minimized during model training to balance the model complexity and data complexity. The final overall loss is taken as the *MDL score* for that causal direction.

4.4.1. Class-Free Implementation

Estimating the Complexity of the Encoding Model

When measuring model simplicity, conventional methods often use regularization terms such as the L1 or L2 norms to encourage the model to use sparser or smaller weight values to ensure simplicity. However, in reality, smaller weight values do not necessarily equate to simplicity. In contrast, we believe that a weight value of 1 should be considered simpler than a smaller value, such as 0.3532, which requires more bits for data transmission. On this basis, we adopted the same weight encoding method used by Nur Lan et al. [18]. This encoding method takes into account the number of bits required to represent the weight values, which aligns better with our simplicity criterion.

$$L(M) = L(\text{parameters}) + L(\text{weights}) \tag{9}$$

The encoding method for the model is specifically described by the following formula. The algorithmic complexity of the model consists of two parts: the hyperparameters

required for model initialization and the weight values of the trained model. The encoding method for these values is based on the prefix-free encoding technique proposed by Li and Vitányi [17], as shown in the following equation:

$$L(\text{numbers}) = \underbrace{11111\ldots111}_{\text{Unary enc. of } \cdot\lceil\log_2 n\rceil}\ \underbrace{0}_{\text{Separator}}\ \underbrace{10101\ldots00110}_{\text{Binary enc. of } \cdot n} \tag{10}$$

Any number is represented as a fraction, where the denominator's maximum length is determined by hyperparameters, which can be understood as the minimum encoding resolution. The numerator is then determined on the basis of the specific value at the resolution provided by the denominator. This encoding method meets the simplicity requirement discussed earlier: intuitively simple numbers such as 1 and 0 require only one bit for the denominator and one bit for the numerator, whereas more complex values require longer bit representations. To ensure fairness in causal direction inference, we use MLP models with the same number of layers and hidden layer sizes for both directions. Therefore, the encoding of the model hyperparameters is identical for both directions, and can be ignored in this step.

Estimate the Complexity of the Encoded Data

In our implementation, the regression model reduces the difference between the modeled conditional distribution and the true conditional distribution by minimizing the regression error. In other words, the regression error reflects the portion of the data's conditional distribution information that the model fails to capture. The smaller the error is, the lower the remaining uncertainty in the data under the model, leading to a shorter conditional description length. Conversely, a larger error indicates an increase in the remaining uncertainty, requiring more bits to model the residual uncertainty information. Naturally, we use the regression error from the regression model to represent the $\mathrm{L}(\text{data} \mid \text{model})$ component in our causal direction inference principle.

4.4.2. Theoretical Analysis of Necessity

The idea of approximating the algorithmic complexity of the conditional distribution via a two-part MDL score from a regression model is not new. In practice, the regression model used to compute the two-part MDL inevitably relaxes the Turing completeness requirement compared with an ideal general Turing machine, which introduces approximation biases that affect the accuracy of causal direction inferences. To discuss this issue more formally, we use the theory of Li and Vitányi [17]. Kolmogorov's structure function, also known as the maximum likelihood estimator (MLE), is defined as follows:

$$h_x(i) = \min_S \{\log|S| : S \ni X, K(S) \leq i\} \tag{11}$$

Here, S is the expected model used to model x, and i is a nonnegative integer that constrains the complexity of the expected model S. We can understand this function as follows: the encoding of x consists of two parts. The first part is the shortest model length i within the predefined expected model space S, and the second part is the information length $h_x(i)$ required to model x under this model. According to the theory presented in this paper, $h_x(i)$ is a nonincreasing function, and the sum $h_x(i) + i$ has the following relationship with $K(x)$:

$$K(x) \leq i + h_x(i) + O(1) \tag{12}$$

Specifically, as i increases, $h_x(i)$ has a lower bound determined by $K(x)$. This lower bound is referred to as the sufficient line $L(i) = K(i) - i$. Additionally, when i exceeds a

certain value α, the model is considered to have captured all the effective information about X. As i increases further, the additional complexity $i - \alpha$ begins to model an increasing amount of noise in the data x, indicating that the model has reached an overfitting state. In this state, $h_x(i)$ closely adheres to the sufficient line. The value $i = \alpha$ is termed the minimal sufficient statistic, at which point $h_x(i)$ first approaches $L(i)$, as illustrated in Figure 3:

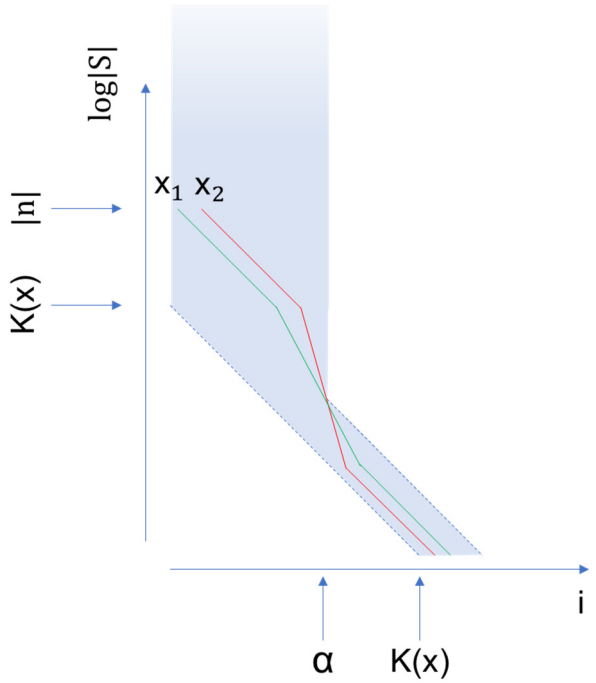


Figure 3. Shape of Kolmogorov's structure function under different data distributions.

In fact, according to this theory, the specific value of α is uncomputable, meaning that we cannot expect the "true" model to be within our model search space. However, our goal is not to find the best model with the best generalization performance but rather to find a metric $h_x(i) + i$ that approximates $K(x)$ for causal direction determination. Therefore, the metric in an overfitting state is as effective as when $i = \alpha$. Additionally, for the same $K(x)$ and $|n|$, different samples x correspond to different shapes of the function $h_x(i)$ within the shaded region. In other words, the deviation of $h_x(i) + i$ from $K(x)$ depends on the specific distribution of the x samples, which could lead to incorrect judgements in the causal direction. Therefore, our key idea is to ensure that the structure functions corresponding to the model have narrower possible shape regions for different x values, thereby reducing the risk of making erroneous judgements.

Intuitively, when instantiating our inference principle, a more general model class should be used to minimize biases caused by model selection as much as possible. For example, when a set of basis functions is used as a predefined search space, if the types of candidate basis functions do not include a specific basis function, the two-part MDL score obtained by approximating with other basis functions will be higher than when the specific basis function is included. In other words, when constructing the predefined model search space, it is necessary to introduce as few priors as possible to ensure that the model does not favor specific data distributions. In terms of causal mechanism functions, this means that we need to fairly assess the complexity of the various forms of the mechanism functions. On the basis of the above considerations, we use an MLP as the regression model, and employ a model weight encoding method as a measure of model simplicity.

5. Experiment Results and Discussion

In this section, we provide a detailed description of the experimental setup, which is specifically divided into two parts: benchmarks and baselines. First, we introduce the different benchmarks used for evaluation, including their selection criteria and relevance to the research objectives. Next, we discuss the selection of the baseline methods, which are used for comparison in experiments to help us better understand the effectiveness of the experimental results and the relative performance of the research methods. Through these introductions, we aim to lay the foundation for the subsequent results analysis and discussion.

5.1. Benchmarks

To validate the effectiveness of our method, we apply it to one widely used real-world dataset and two synthetic datasets. The real-world data reflect the complex causal structures present in practical problems, including potential noise, confounding factors, and nonlinear relationships. Therefore, we can use the method's performance on these datasets as a direct measure of its effectiveness in real-world applications. Additionally, synthetic data provide a controlled environment that can simulate specific causal generation mechanisms. We utilize them to verify the stability of our method under different data generation mechanisms, such as the presence or absence of noise and potential confounding factors.

Real-world data: The Tübingen cause–effect pairs dataset (TCEP) dataset was released by researchers from the University of Tübingen in Germany. Each sample contains two variables and clearly indicates the direction of the causal relationship between them. The dataset is collected from multiple other datasets across different professional fields, encompassing data samples from disciplines such as physics, economics, biology, and social sciences. Owing to the diversity of samples and precise annotations, it has become a widely used benchmark dataset in the field of causal inference. At the time of writing, the dataset contains a total of 108 cause–effect pairs, and is still being continuously expanded. Therefore, in the baseline models discussed later, the accuracy of some models may differ from the accuracy reported in their original papers. To ensure fairness, for methods with significant differences, we referenced the baseline model accuracies reported in a recent review on causal direction determination [32].

Synthetic data: The synthetic datasets we selected come from two widely used sources. Four synthetic causal pairs examine different noise distributions and confounder scenarios, namely, SIM (without confounders), SIM-ln (with low noise), SIM-G (with distributions close to Gaussian), and SIM-c (with confounders). Additionally, three synthetic causal pairs investigate different noise mechanisms, including additive noise (AN), nonlinear noise (LS), and multiplicative uniform noise (MN-U). Each scenario in these synthetic datasets includes a certain number of mutually independent samples to assess the performance of the methods under these settings.

5.2. Baselines

In selecting baseline models, we considered both the ICM-based methods that adhere to the Solomonoff prior, and the other classic methods that do not follow the Solomonoff prior. Among the ICM-based methods, we select a representative and state-of-the-art series of algorithms, including the CURE, IGCI [33], QCCD [34], and slope algorithms. Other classic algorithms include Lingam, ANM, PNL, and RESIT [35]. Additionally, we selected a recently emerged method, CEMM, as a representative of the state-of-the-art functional causal model-based approaches.

5.3. Results

5.3.1. Real-World Data

In the experimental setup of our proposed method, certain hyperparameters need to be configured (e.g., the architecture of the backbone network). To avoid excessive hyperparameter tuning, and ensure the objectivity of the experimental results, we used the same hyperparameter settings for each of the 108 data pairs included in TCEP. Except for a few pairs where one or two variables (cause or effect) are multivariate, we adjust some necessary parameters (such as the sizes of the input and output layers) to ensure that the data can be properly input into the model. The performance of each method on the TCEP dataset is shown in Figure 4.

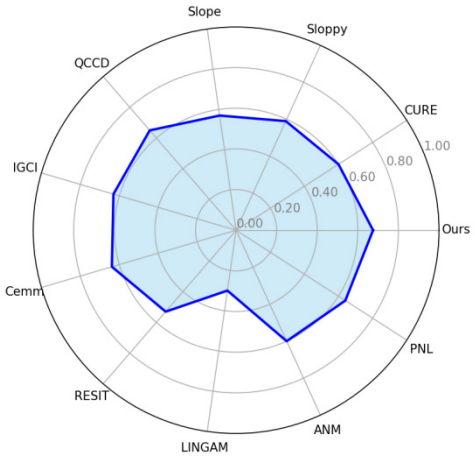


Figure 4. Comparison of method performances on the TCEP dataset.

Since real-world data reflect causal mechanisms, noise mechanisms, and other information that are generally considered uncontrollable and diverse in real-world usage scenarios, the chosen methods need to use the most general and robust simple assumptions possible when designing causal direction inference principles. One of the objectives of our proposed method is to explore such simple and general assumptions for causal direction inference tasks. In Figure 4, the upper half shows the ICM-based methods that follow the Solomonoff prior, as well as our proposed NMDL4CO method inspired by the INUS condition on the right. As theoretically expected, Figure 4 shows that these methods exhibit excellent performance with relatively small differences, with one reason being that the methods adhering to the Solomonoff prior typically do not make explicit restrictive assumptions about the functional forms of data generation mechanisms or noise distributions. This allows such methods to perform better when dealing with real-world datasets that contain various complex causal relationships. In contrast, the methods in the lower half of Figure 4 require explicit assumptions about the data generation mechanisms, resulting in greater performance variability on real-world datasets. Although the CEMM method, owing to its unique design, is no longer limited to a single assumption, its effectiveness theoretically still depends on the data generation mechanism conforming to at least one of the existing assumptions. Furthermore, compared with other methods that also follow the Solomonoff prior, our method achieves an accuracy of 68.5%, demonstrating competitive performance. The superior performance of our method over the other methods also indicates that the theoretical foundation of our method does not rely on the assumption that "the effect variable does not contain other unobserved causes", making it more consistent with real-world scenarios. The relatively small performance differences are attributed, in our view, to a considerable portion of real-world data satisfying the assumption that the effect variable does not contain other unobserved cause variables. In other words, in our

observed samples, there exist cases where other unobserved cause variables are either consistent or vary minimally, thereby having little impact on the effect. For example, certain factors influenced by geographical location remain largely consistent because of sampling within the same area. For example, in the TCEP dataset, pairs 49–51 record the impact of temperature from three regions on near-ground ozone concentrations, and pairs 94–96 record the relationships among the number of hours in a day, temperature, and hourly electricity consumption at a certain location. Since the chemical reaction activity for ozone generation is influenced primarily by temperature, temperature is the key factor affecting the near-ground ozone concentration. Additionally, the near-ground ozone concentration is also influenced by local microclimatic conditions, such as urban inversions that cause ozone to accumulate in the environment. Since these three data pairs are collected in fixed regions, we can simply assume that the influence of these factors on the near-ground ozone concentration is rarely reflected in the data samples. Similarly, residents' electricity consumption at different times is influenced by their daily routines, which are in turn regulated by the solar altitude angle. The solar altitude angle can be approximated by the number of hours in a day; in the absence of drastic weather changes, the temperature within a day can also reflect changes in the solar altitude angle. On this basis, within specific time periods, the temperature and hour counts can be regarded as the primary influencing factors on residents' electricity consumption, because other potential influencing factors (such as regional economic levels) typically do not undergo significant changes during short data collection periods. The causal independence mechanism assumption is more applicable to these data pairs. We compare the performance of our method with that of the classical SLOPE and IGCI methods, which represent the two most representative ideas based on the Solomonoff prior: (1) utilizing the relationship between the algorithmic complexities of marginal and conditional distributions and (2) leveraging the independence between the cause and the causal mechanism. The results are shown in Table 2, Symbol ✓ represents a correct method identification result, while symbol ✘ represents an incorrect method identification result.

Table 2. Comparison of method performances under conditions where the impact of unobserved cause variables is not significant.

	Ours (0%)	Slope (100%)	IGCI (83%)
Pair 49	✘	✓	✓
Pair 50	✘	✓	✓
Pair 51	✘	✓	✘
Pair 94	✘	✓	✓
Pair 95	✘	✓	✓
Pair 96	✘	✓	✓

In contrast, when there are clearly other unobserved cause variables present, such as in pairs 43–46 and 56–63 of the TCEP dataset, our method performs significantly better. Pairs 43–46 describe the impacts of temperature, ground pressure, sea level pressure, and relative humidity at the previous time step on the next time step within a certain range. Pairs 56–63 describe the influence of capital latitude on the life expectancy of males and females. The performance on these datasets is shown in Table 3.

Table 3. Comparison of method performances under conditions of significant impact from unobserved cause variables.

	Ours (91%)	Slope (25%)	IGCI (58%)
Pair 43	✓	✘	✓
Pair 44	✓	✘	✘
Pair 45	✘	✓	✘

Table 3. *Cont.*

	Ours (91%)	Slope (25%)	IGCI (58%)
Pair 46	✓	✘	✓
Pair 56	✓	✘	✘
Pair 57	✓	✘	✓
Pair 58	✓	✓	✓
Pair 59	✓	✘	✓
Pair 60	✓	✘	✘
Pair 61	✓	✘	✘
Pair 62	✓	✓	✓
Pair 63	✓	✘	✓

5.3.2. Synthetic Data

Figure 5 visualized the performance of Solomonoff prior-based methods on synthetic data using radar charts. Under all the controlled environmental settings in the synthetic data, our proposed method consistently demonstrated a relatively stable and reliable performance, maintaining a top-tier performance in most settings. In other words, our method is capable of handling the majority of cases involving complex data generation mechanisms, including various functional forms and noise distributions or combinations, when facing real-world application scenarios.

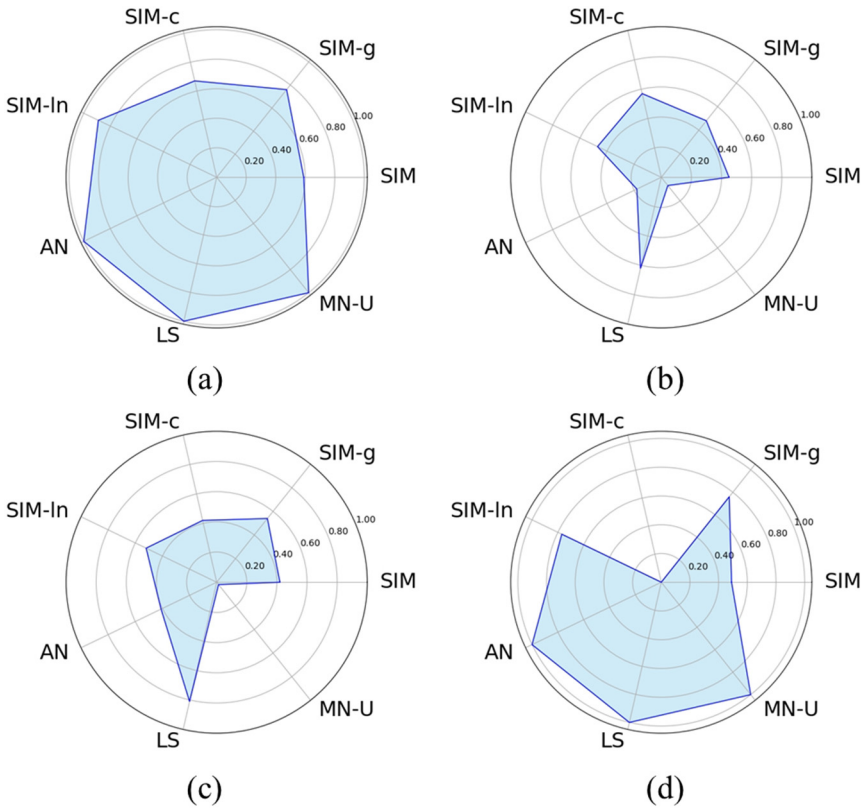


Figure 5. Comparison of the performance of Solomonoff prior-based methods on synthetic data. (**a**) Our method, (**b**) SLOPE, (**c**) IGCI, and (**d**) QCCD are included. The QCCD method does not conform to the assumptions of SIM-c, and is therefore excluded from the comparison. Additionally, the CURE method is not included in the comparison because its code is unavailable.

This finding indicates that our analysis was effective when instantiating the concept of Kolmogorov complexity involved in our inference principle. When designing causal inference methods based on the Solomonoff prior, it is essential to avoid imposing explicit restrictions on the model class when approximating Kolmogorov complexity. On

the one hand, this helps to prevent the method from exhibiting sensitivity to data generation mechanisms, similar to FCM-based methods. On the other hand, preserving Turing completeness better ensures a more accurate approximation of Kolmogorov complexity, thereby enhancing the method's performance. For example, the QCCD method, which similarly does not impose explicit restrictions on the model class, not only performs excellently in terms of sensitivity to data generation mechanisms, but also exhibits a relatively superior overall performance. In contrast, the SLOPE method, which is very similar to our method, shows significant performance discrepancies in many cases, especially under the multiplicative uniform noise (MN-U) setting, where the method almost completely fails. Additionally, the IGCI method performs poorly under the SIM-ln setting, which contradicts its declared requirement for low-noise environments. One explanation for this, is that under the SIM-ln setting, the causal mechanism functions exhibit more non-injective or near-linear relationships, which is also detrimental to the IGCI's causal direction determination.

This also reminds us to be cautious when using synthetic data to evaluate methods' performances under specific environments, because during the synthesis process, certain technical reasons might lead us to not only restrict the conditions we intend but also inadvertently alter some conditions that should remain unchanged. This is why we prefer to demonstrate the validity of our method's theoretical assumption through its performance on real-world data rather than by constructing a specific synthetic dataset. The overall performance of the methods on the synthetic datasets is shown in Table 4. The bold font represents the performance of the SOTA method on this data.

Table 4. Comparison of method performances on synthetic data.

	sim	**sim-g**	**sim-c**	**sim-ln**	**AN**	**LS**	**MN-U**
ours	0.59	0.76	**0.67**	0.84	**1**	**1**	**1**
slope	0.45	0.48	0.57	0.47	0.18	0.62	0.07
qccd	0.49	0.76		0.77	1	1	1
igci	0.42	0.54	0.42	0.52	0.41	0.81	0.02
cemm	0.7	**0.88**	0.65	0.79	0.98	0.89	0.85
resit	**0.78**	0.77		**0.87**	1	0.31	0.09
lingam	0.44	0.33	0.47	0.32	0.04	0	0
anm	0.39	0.39	0.35	0.38	1	0.55	0.21
pnl	0.19	0.12	0.23	0.16	0.95	0.97	0.96

6. Conclusions and Future Work

The main contribution of this study is the introduction of a new hypothesis, inspired by the INUS condition, to the field of causal inference, based on the Solomonoff prior, and the design of a new causal inference principle on the basis of this hypothesis. Compared with the original ICM assumption, our proposed hypothesis aims to overcome its counter-intuitive assumption that the effect in causal pairs does not include other unobserved cause variables. In contrast, we believe that other unobserved cause variables are inherently present, and play a crucial role in our causal direction determination. Additionally, we analyzed how the approximation method of Kolmogorov complexity affects the effectiveness of designing causal direction determination methods based on the Solomonoff prior. On the basis of this understanding, we developed a specific algorithm on NMDL. The experimental results on widely used real-world and synthetic datasets demonstrate that our method performs excellently and remains stable.

However, it is important to note that we do not consider that our hypothesis fully replaces the ICM assumption. Although we intuitively believe that the set of potential cause variables for the effect variable is comprehensively observed in practical applications, it is undeniable that causal mechanisms can be relatively simple, making comprehensive

observation possible. From real-world experimental results, it can be seen that in pairs 49–51 and 94–96, where we can assume that no unobserved variables significantly affect the outcome variable, our method's accuracy was 0%, showing a clear failure. In contrast, the two ICM-based methods achieved accuracies of 100% and 83%, respectively. A more reasonable approach is to balance simple, comprehensive observations and complex, incomprehensive observations. Therefore, exploring unified hypotheses and causal direction determination principles that bridge these two extremes is a potential future direction. Furthermore, developing more accurate approximation methods for Kolmogorov complexity is an effective means to enhance the performance of causal direction determination methods based on the Solomonoff prior, and is equally worthy of further exploration.

Author Contributions: Conceptualization, Z.C. and Q.L.; methodology, Z.C.; formal analysis, Z.C.; investigation, Z.C.; writing—original draft, Z.C.; writing—review and editing, S.H.; supervision, L.Z., Q.L. and S.H.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 42171458, the National Natural Science Foundation of China under Grant 42271481 and the Natural Science Foundation of Hunan Province under Grant 2022JJ30698.

Data Availability Statement: Datasets used in this study are available at https://webdav.tuebingen.mpg.de/cause-effect/ (accessed on 14 November 2023) and https://github.com/tagas/qccd/tree/master/data/ (accessed on 27 June 2024).

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

References

1. Schölkopf, B.; von Kügelgen, J. From Statistical to Causal Learning. *arXiv* **2022**, arXiv:2204.00607. [CrossRef]
2. Camps-Valls, G.; Gerhardus, A.; Ninad, U.; Varando, G.; Martius, G.; Balaguer-Ballester, E.; Vinuesa, R.; Diaz, E.; Zanna, L.; Runge, J. Discovering Causal Relations and Equations from Data. *Phys. Rep.* **2023**, *1044*, 1–68. [CrossRef]
3. Imbens, G.W.; Rubin, D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: Cambridge, MA, USA, 2015.
4. Scheffer, M.; Brovkin, V.; Cox, P.M. Positive Feedback between Global Warming and Atmospheric CO_2 Concentration Inferred from Past Climate Change. *Geophys. Res. Lett.* **2006**, *33*, 2005GL025044. [CrossRef]
5. Watanabe, C. The Feedback Loop between Technology and Economic Development: An Examination of Japanese Industry. *Technol. Forecast. Soc. Change* **1995**, *49*, 127–145. [CrossRef]
6. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*, 1st ed.; Basic Books, Inc.: New York, NY, USA, 2018; ISBN 0-465-09760-X.
7. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: Cambridge, MA, USA, 2009; ISBN 0-521-89560-X.
8. Lemeire, J.; Janzing, D. Replacing Causal Faithfulness with Algorithmic Independence of Conditionals. *Minds Mach.* **2013**, *23*, 227–249. [CrossRef]
9. Peters, J.; Mooij, J.; Janzing, D.; Schoelkopf, B. Identifiability of Causal Graphs Using Functional Models. *arXiv* **2012**, arXiv:1202.3757. [CrossRef]
10. Janzing, D.; Schoelkopf, B. Causal Inference Using the Algorithmic Markov Condition. *arXiv* **2008**, arXiv:0804.3678. [CrossRef]
11. Vonk, M.C.; Malekovic, N.; Bäck, T.; Kononova, A.V. Disentangling Causality: Assumptions in Causal Discovery and Inference. *Artif. Intell. Rev.* **2023**, *56*, 10613–10649. [CrossRef]
12. Rathmanner, S.; Hutter, M. A Philosophical Treatise of Universal Induction. *Entropy* **2011**, *13*, 1076–1136. [CrossRef]
13. Marx, A.; Vreeken, J. Telling Cause from Effect by Local and Global Regression. *Knowl. Inf. Syst.* **2019**, *60*, 1277–1305. [CrossRef]
14. Hoyer, P.O.; Janzing, D.; Mooij, J.; Peters, J.; Schölkopf, B. Nonlinear Causal Discovery with Additive Noise Models. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–10 December 2008; Curran Associates Inc.: Red Hook, NY, USA, 2008; pp. 689–696.
15. Janzing, D. The Cause-Effect Problem: Motivation, Ideas, and Popular Misconceptions. In *Cause Effect Pairs in Machine Learning*; Guyon, I., Statnikov, A., Batu, B.B., Eds.; The Springer Series on Challenges in Machine Learning; Springer International Publishing: Cham, Switzerland, 2019; pp. 3–26, ISBN 978-3-030-21809-6.

16. Denise, T.C. On the Nature of INUS Conditionality. *Analysis* **1984**, *44*, 49–52. [CrossRef]
17. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Texts in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; ISBN 978-3-030-11297-4.
18. Lan, N.; Chemla, E.; Katzir, R. Bridging the Empirical-Theoretical Gap in Neural Network Formal Language Learning Using Minimum Description Length. *arXiv* **2024**, arXiv:2402.10013. [CrossRef]
19. Wu, J.; Zhou, Y.; Wang, H.; Wang, X.; Wang, J. Assessing the Causal Effects of Climate Change on Vegetation Dynamics in Northeast China Using Convergence Cross-Mapping. *IEEE Access* **2023**, *11*, 115367–115379. [CrossRef]
20. Zhao, Y.; Ma, E.; Zhou, Z.; Zou, Y.; Cao, Z.; Cai, H.; Li, C.; Yan, Y.; Chen, Y. Detecting the Non-Separable Causality in Soil Moisture-Precipitation Coupling With Convergent Cross-Mapping. *Water Resour. Res.* **2024**, *60*, e2023WR034616. [CrossRef]
21. Stam, C.J. Nonlinear Dynamical Analysis of EEG and MEG: Review of an Emerging Field. *Clin. Neurophysiol.* **2005**, *116*, 2266–2301. [CrossRef]
22. Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; Fox, E. Neural Granger Causality. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4267–4279. [CrossRef] [PubMed]
23. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.; Deyle, E.; Fogarty, M.; Munch, S. Detecting Causality in Complex Ecosystems. *Science* **2012**, *338*, 496–500. [CrossRef] [PubMed]
24. Butler, K.; Feng, G.; Djurić, P.M. On Causal Discovery With Convergent Cross Mapping. *IEEE Trans. Signal Process.* **2023**, *71*, 2595–2607. [CrossRef]
25. Gao, B.; Yang, J.; Chen, Z.; Sugihara, G.; Li, M.; Stein, A.; Kwan, M.-P.; Wang, J. Causal Inference from Cross-Sectional Earth System Data with Geographical Convergent Cross Mapping. *Nat. Commun.* **2023**, *14*, 5875. [CrossRef] [PubMed]
26. Shimizu, S.; Hoyer, P.O.; Hyvärinen, A.; Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.* **2006**, *7*, 2003–2030.
27. Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; Sebag, M. Learning Functional Causal Models with Generative Neural Networks. *arXiv* **2017**, arXiv:1709.05321. [CrossRef]
28. Varando, G.; Catsis, S.; Diaz, E.; Camps-Valls, G. Pairwise Causal Discovery with Support Measure Machines. *Appl. Soft Comput.* **2024**, *150*, 111030. [CrossRef]
29. Budhathoki, K.; Vreeken, J. Origo: Causal Inference by Compression. *Knowl. Inf. Syst.* **2018**, *56*, 285–307. [CrossRef]
30. Sgouritsa, E.; Janzing, D.; Hennig, P.; Schölkopf, B. Inference of Cause and Effect with Unsupervised Inverse Regression. In Proceedings of the UAI 2015 Conference on Advances in Causal Inference, Amsterdam, The Netherlands, 16 July 2015; CEUR-WS.org: Aachen, Germany, 2015; Volume 1504, p. 89.
31. Mooij, J.M.; Peters, J.; Janzing, D.; Zscheischler, J.; Schölkopf, B. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *J. Mach. Learn. Res.* **2016**, *17*, 1103–1204.
32. Käding, C.; Runge, J. Distinguishing Cause and Effect in Bivariate Structural Causal Models: A Systematic Investigation. *J. Mach. Learn. Res.* **2024**, *24*, 13077–13220.
33. Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniušis, P.; Steudel, B.; Schölkopf, B. Information-Geometric Approach to Inferring Causal Directions. *Artif. Intell.* **2012**, *182–183*, 1–31. [CrossRef]
34. Tagasovska, N.; Chavez-Demoulin, V.; Vatter, T. Distinguishing Cause from Effect Using Quantiles: Bivariate Quantile Causal Discovery. *arXiv* **2018**, arXiv:1801.10579. [CrossRef]
35. Peters, J.; Mooij, J.; Janzing, D.; Schölkopf, B. Causal Discovery with Continuous Additive Noise Models. *arXiv* **2013**, arXiv:1309.6779. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.