entropy

MDPI

Article

Neural Causal Information Extractor for Unobserved Causes

Keng-Hou Leong [1,2], Yuxuan Xiu [1,2], Bokui Chen [1,3,*] and Wai Kin (Victor) Chan [1,2,4,*]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China; liangjh22@mails.tsinghua.edu.cn (K.-H.L.); xiuyx19@mails.tsinghua.edu.cn (Y.X.)
[2] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China
[3] Peng Cheng Laboratory, Shenzhen 518055, China
[4] International Science and Technology Information Center, Shenzhen 518055, China
* Correspondence: chen.bokui@sz.tsinghua.edu.cn (B.C.); chanw@sz.tsinghua.edu.cn (W.K.C.)

Abstract: Causal inference aims to faithfully depict the causal relationships between given variables. However, in many practical systems, variables are often partially observed, and some unobserved variables could carry significant information and induce causal effects on a target. Identifying these unobserved causes remains a challenge, and existing works have not considered extracting the unobserved causes while retaining the causes that have already been observed and included. In this work, we aim to construct the implicit variables with a generator–discriminator framework named the Neural Causal Information Extractor (NCIE), which can complement the information of unobserved causes and thus provide a complete set of causes with both observed causes and the representations of unobserved causes. By maximizing the mutual information between the targets and the union of observed causes and implicit variables, the implicit variables we generate could complement the information that the unobserved causes should have provided. The synthetic experiments show that the implicit variables preserve the information and dynamics of the unobserved causes. In addition, extensive real-world time series prediction tasks show improved precision after introducing implicit variables, thus indicating their causality to the targets.

Keywords: causal inference; maximizing mutual information; unobserved causes; complex system

Citation: Leong, K.-H.; Xiu, Y.; Chen, B.; Chan, W.K. Neural Causal Information Extractor for Unobserved Causes. *Entropy* **2024**, *26*, 46. https://doi.org/10.3390/e26010046

Academic Editors: Jiang Zhang, Peng Cui and Hector Zenil

Received: 6 November 2023
Revised: 18 December 2023
Accepted: 22 December 2023
Published: 31 December 2023

1. Introduction

Identifying the complete and valid causes of a variable has been a significant pursuit in causal inference toward complex systems. A causal inference approach is called "causally faithful" [1] when it reveals the causal relationships shown in the observed variables, and it is "causally sufficient" [2] if it does not omit any of the hidden confounder or unobserved variables that should be included. To hold causal sufficiency when deriving faithful causal relationships, traditional causality inference methods [1–6] often assume the complete observability of the system, i.e., all of the variables and information within the system are collected and observed. In the perspective of [7,8], the information within the system flows from the causes to the targets; thus, the complete causes can provide most of the target's information and can describe it (except the noise terms) in detail, as shown in Figure 1a.

However, in practical scenarios, complex systems are often partially observable since only a subset of variables can be directly measured. For instance, in financial markets, the influence of media [9] and company-specific factors [10], which are hardly measurable, drive important causal effects in local tradings [9] and price reversals [10]. Another example can be shown in ecosystems, in which the dynamics are described by species populations, habitat quality, and threats; however, one can rarely acquire a perfect knowledge of the system states [11,12] to perform adequate management actions. The partial observability also induces analyzing challenges in domains such as monitoring biophysical objects [13], weather and climate forecasting [14], and industrial management [15]. While part of the causes are unobserved, the target's information that should be provided is unknown,

as shown in Figure 1b. To describe the causal structures in such incomplete systems, causal sufficiency in traditional works is no longer guaranteed, and their results do not necessarily identify the real causes even when they are causally faithful [1,16].

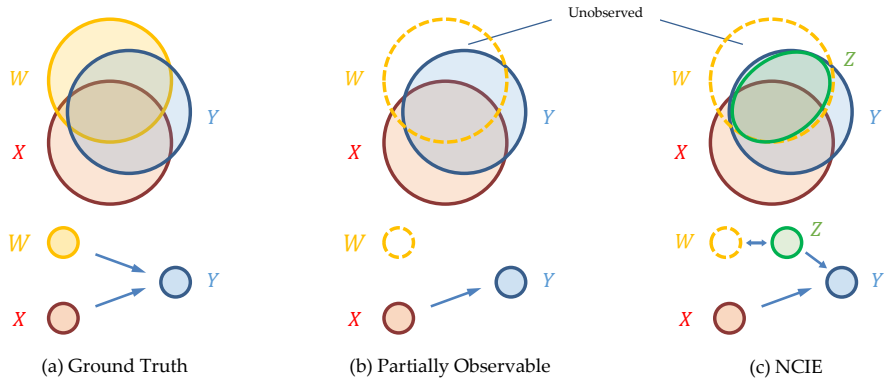

Figure 1. Illustration of the information in a Venn diagram, as well as the causal relationships of the (**a**) ground truth; (**b**) partially observable scenario in reality; and (**c**) an alternative representation of causal relationships with the implicit variables Z, which is generated by NCIE, where Y, X, and W denote the target, observed causes, and unobserved causes, respectively.

Therefore, to avoid identifying spurious causes, the common idea shared by most practices [17–28] is to introduce implicit variables that compensate for the missing causes. Most of them have applied neural networks as a powerful tool for generating variables, which could be categorized into two streams: either (i) employing neural networks with latent layers to uncover the significant features in observed variables that better formulate the targets, or (ii) utilizing the ability to preserve information in representation learning to extract unobserved or latent causes that provide significant information of the targets.

The works from the first approach adopt observed causes as the inputs to train a neural network that minimizes the fitting loss between the output and the target. The intermediate layers in the neural networks help to better fit the target in supervised learning [17], for example, by capturing the hidden long-/short-term dependencies in the dynamics [18,19]. Granger [29] suggested that we could consider these hidden variables as causes, since including them could perform a better regression of the targets. However, such methods are, in actuality, just recombining the neurons from the input to extract advanced features, and they cannot provide information beyond the input [30]. Therefore, they do not necessarily provide unobserved information; thus, we can hardly identify these hidden variables as those that are unobserved in the system.

The second stream aims to generate implicit variables that provide more information to the targets. This echoes another definition of causality [3,31], in which a cause should conditionally correlate with the target given any of the sets of other variables, or it could provide information that all the other variables could not provide to the target [32]. These methods [20–28] adopt the idea of representation learning, which inputs the target (and all the observed variables) into neural networks to project an embedding. This preserves the information while describing the prominent dynamics of the original data. Since these latent variables generate a cover of most of the targets' information, they could be referred to as the causes that drive the system. However, in general, the observed and unobserved causes are entangled and inseparable in the causal structure. While refs. [20–28] excluded all of the observed causes and only considered the generated latent variables as the targets' causes, their depiction does not necessarily reveal the real causal relationships.

Motivated by the aforementioned issues of either missing the unobserved information or giving up on the observed causes, this paper proposes a framework to complete the causal structure by extracting the unobserved variables while retaining the observed causes. We assume the existence of unobserved causes to the target variables; hence, the targets' information should be provided by both the observed and unobserved causes. Given the

part of the information from the observed causes, our objective is to generate implicit variables from the observed variables that provide as much information to the target as possible, thus covering the information that should have been provided by the unobserved causes. As shown in Figure 1c, the implicit variables Z would cover the information of the target Y that should have been provided by the unobserved causes W, and hence it could provide an alternative representation ($\{X, Z\} \rightarrow Y$) of the information flow in the causal relationships $\{X, W\} \rightarrow Y$.

In detail, we employ a generator–discriminator architecture named the Neural Causal Information Extractor (NCIE) that extracts and embeds the targets' information into the implicit variables. The generator obtains implicit variables fromthe time series inputs, and the discriminator maximizes the mutual information between the targets and the union of explicit and implicit causes by constraining implicit variables to have diverse marginal and joint distributions. This results in a holistic causal structure that encompasses both observed and unobserved factors to the target. To highlight the ability of NCIE to complement unobserved information and recover unobserved dynamics, we generate implicit variables for three synthetic cases. To further demonstrate the efficacy of the generated implicit causes in improving time series predictions and to verify their causality, we perform extensive experiments on real-world time series data.

The contribution of our work can be highlighted as follows:

1. We propose a generator–discriminator architecture that generates implicit variables to complement the unobserved causes in the causal structures while retaining the observed causes.
2. The implicit variables we generate could carry information from the unobserved variables and reveal their dynamics.
3. Time series prediction tasks show that the combination of observed and implicit variables helps improve the prediction of targets, verifying that they are better candidates for causal inference.

This paper is organized as follows. Section 2 provides a comprehensive review of related works. Section 3 presents the basic theories in information theory and causal inference and states the objective of this work. Section 4 introduces the methodology of the proposed approach. Section 5 details the experimental setup and dataset descriptions, and then presents the results and discussion of our experiments. Finally, in Section 6, we provide a discussion on the implications of our work and present future possible advanced works.

2. Literature Review

In the scope of our work, we categorize the related works on causal inference into two parts, which cover the methods with and without considering unobserved variables.

2.1. Causal Inference without Considering Unobserved Variables

Assuming that the given observations are complete and thus sufficient, the objectives of traditional works often include drawing causal faithful graphs, in which nodes represent the variables of interest and edges indicate the existence of causal relationships between variables. The constrained-based methods, including the Peter–Clark (PC) algorithm [2] and its improvements [2,3,33], are graph algorithms that start with a fully connected graph and iteratively delete the edges between two variables if they are conditionally independent or d-separated [2]. They provide causal faithful graphs that satisfy the causal Markov condition [31], i.e., a variable is conditionally independent to all variables except its effects given its causes. On the other hand, score-based approaches present the causal structure with the Bayesian network with the highest score, measured by the Bayesian Dirichlet equivalent (BDe) score [4] and the K2 score [5]. Though the latter methods do not ensure faithfulness, they pursue a representation that shows the causal structure and, at the same time, the strength of cause-to-effect dependencies. Other approaches include those

based on Granger causality [6], which also aim to identify the existence and magnitudes of the causal edges among variables.

2.2. Causal Inference Considering Unobserved Variables

There are plenty of works on causal inference that are aware of the existence of unobserved variables, and for clarity, we attempt to classify them into three groups. The first group refers to those that detect the existence of hidden confounders without knowing the effects, including Fast Casual Inference (FCI) [2] and Latent PCMCI (LPCMCI, where PCMCI [3] is a PC-based condition selection with the Momentary Conditional Independence (MCI) test) [34]. They are the modified versions of the PC algorithm [2] and PCMCI [3], respectively, which follow similar procedures but identify an extra kind of edge (relationship) between variables: bidirectional arrows. This indicates the conditional correlational relationship between two variables that could not be explained by any directional causal path from one to another, thus identifying the existence of a hidden confounder driving the two variables.

The second group attempts to explain the causal effects originating from unobserved variables toward the target variable. Assuming linear, additive, and first-order Markovian causal effects, ref. [16] constructs a linear regression model that describes the observed variables by past observed variables, unobserved causes, and noises. Based on this model, the effects of the unobserved causes, which refer to their coefficients, could be obtained by variational expectation maximization or by solving the autocovariance.

The third group [20–28] aims to identify the unobserved causes via latent representations. They assume that the observed signals are driven or caused by a set of unobserved latent signals, and there exists a latent dynamic that drives the latent variables from past to future. Therefore, given the observed dynamics, their objectives are to identify the posterior distributions of future latent signals given past latent signals, the posterior distributions of observed signals given the contemporaneous latent signals, and the latent signals themselves. They identify latent signals via different approaches, including Variational Autoencoder (VAE) [20,21,35,36], Dynamical Component Analysis [25–27], and maximizing the (conditional) mutual information [22,28].

While refs. [20–28] ensure the identifiability of implicit variables, their structures do not include the causes that are already observed, which might also be involved in the causal process. In our work, we express the causal structures containing both observed and implicit causes. Moreover, while refs. [20–28] could precisely identify the implicit variables, many of their generators, such as VAE [20,21,35,36], hold specific assumptions about the implicit variables. With a simple generator that holds comparably few assumptions about the distribution, our architecture can be applied to more general cases.

2.3. Mutual Information Estimator

$$\begin{aligned} H(Y) &= -\sum_{y \in Y} p(y) log(p(y)), \\ H(Y) &= -\int_{Y} p(y) log(p(y)) dy. \end{aligned} \tag{1}$$

In Equation (1), Shannon provided an information theory-based definition for the entropy of random variables with known distributions. However, the distribution of variables in real-world time series remains unknown in many scenarios. To sample the unknown distributions, one of the fundamental methods is binning [37,38]. By partitioning the continuous variables into discrete bins and counting the sampled distributions, one can

apply Equation (1) to estimate entropy. Provided the entropy of variables, one can calculate the mutual information between two variables, X and Y, with

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right) \\ &= H(X)+H(Y)-H(X,Y). \end{aligned} \tag{2}$$

The Mutual Information Neural Estimator (MINE) [39] is another popular method that uses a neural network to estimate the KL Divergence of two random variables, which could then provide a tight lower bound for their mutual information. Deep InfoMax [22] provides a similar approach by estimating Jensen–Shannon (JS) Divergence. Under the Gaussian assumption, Refs. [25,26] measure entropy by $H(Z) = \frac{r1}{2} ln(2\pi e)^{dT}|\Sigma_T(Z)|$, and prove that the estimation is still tight if the assumption is violated. Refs. [40–42] provide other effective estimators for information based on classifiers and k-nearest neighbor (kNN).

3. Problem Statement

Suppose we have target variable Y_t at time t. First, we consider the case where Y_t is auto-correlated, and all the causes of Y_t are known, denoted by $\{X_t, Y_{<t}\}$, where $X_t = \{X^1, X^2, \dots, X^n\}_t$ and $Y_{<t} = \{Y_1, \dots, Y_{t-1}\}$. In this work, we assume a k-order Markovian property on auto-correlated causality, i.e., $Y_{<t} = \{Y_{t-k}, \dots, Y_{t-1}\}$.

For simplicity, we abbreviate Y_t as Y. While all the causes $\{X, Y_{<t}\}$ are known, we assume

$$H(Y) = I(Y;\ X,\ Y_{<t}), \tag{3}$$

where $H(Y)$ is the entropy of Y, a measure of uncertainty, and $I(Y;\ X,\ Y_{<t})$ is the mutual information (MI) between Y and $\{X,\ Y_{<t}\}$. MI is a measure of "information" or "reduction of entropy" that the two variables provide to each other. In Equation (3), the entropy reduction that X and $Y_{<t}$ provide to Y equals the entropy of Y. Hence, the causes of Y could eliminate all the uncertainty of Y, in other words, provide all the information of Y.

Now, we consider another case in which, besides some observed causes $X = \{X^1,\ X^2,\ \dots,\ X^n\}$, there are other unobserved causes W. Therefore, we assume

$$H(Y) = I(Y;\ X,\ Y_{<t},\ W). \tag{4}$$

The absence of information about W induces

$$H(Y) > I(Y;\ X,\ Y_{<t}). \tag{5}$$

As shown in Figure 2, X and $Y_{<t}$ could only provide a part (the red part) of the information to Y. Therefore, our objective is to generate some implicit variables $Z = \{Z^1,\ Z^2,\ \dots,\ Z^m\}$ to maximize $I(Y;\ X,\ Y_{<t},\ Z)$ (the yellow part), which attempts to approach its upper bound $H(Y)$ (the large black box).

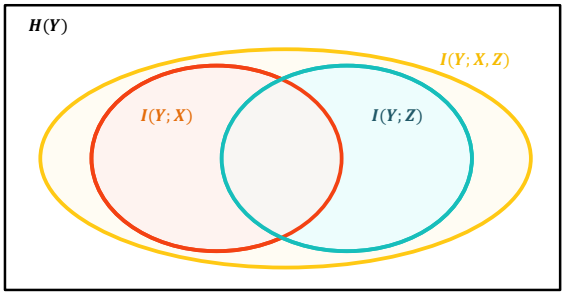

Figure 2. The Venn Diagram illustration of the information provided to Y by X (red), Z (blue), and $\{X, Z\}$ (yellow).

Hence, if we could search for a Z that could perfectly provide this information, we could say that Z somehow represents the unobserved causes W. More specifically, Z traces the information for Y from W.

While the upper bound of the MI equals the entropy of Y, i.e.,

$$\sup_{Z} I(Y;\ X,\ Y_{<t},\ Z) = H(Y), \tag{6}$$

the objective of our work is to generate Z which maximizes the MI and makes it approach $H(Y)$:

$$\max_{Z} I(Y;\ X,\ Y_{<t},\ Z), \tag{7}$$

which is equivalent to maximizing the conditional mutual information of Z and Y given X and $Y_{<t}$:

$$I(Y;Z|X,\ Y_{<t}) = I(Y;\ X,\ Y_{<t},\ Z) - I(Y;\ X,\ Y_{<t}). \tag{8}$$

Since the term without Z could be considered constant here,

$$\max_{Z} I(Y;\ X,\ Y_{<t},\ Z) = \max_{Z} I(Y;Z|X,\ Y_{<t}). \tag{9}$$

Objective (7) is maximized if Z covers all the conditional mutual information from W,

$$\sup_{Z} I(Y;\ Z|\ X,\ Y_{<t}) = I(Y;\ W|\ X,\ Y_{<t}) = H(Y|X), \tag{10}$$

and meanwhile $X, Y_{<t}, Z$ together cover all the entropy (Equation (6)) and provide all the information of Y.

There are some assumptions made on Z:

1. We focus on the unobserved variables that provide information to Y. Those without information of Y do not affect the distributions of Y.
2. In the case of time series, we denote $X_{<t}$, $Z_{<t}$ as the explicit and implicit causes of Y_t to uphold the assumption of temporal priority [1], i.e., causal relationships only move from past variables to future variables in time series. For simplicity, we denote $X_{<t}$ and $Z_{<t}$ as X and Z, and Y_t as Y.
3. We assume X_t, Y_t, and Z_t are in a closed system and there is a causal loop among them, i.e., $X_{<t}$ and $Z_{<t}$ may induce Y_t, and $X_{<t}$ and $Y_{<t}$ may induce Z_t.
4. While we are generating $Z_{<t}$, we do not necessarily obey temporal priority [1], which is a common practice in refs. [20–28,30]. Therefore, we employ $X_{<t}$ and Y_t to generate $Z_{<t}$. We note that we do not focus on finding the exact causes of Z here. Instead of finding causes of Z by applying the whole structures we use on Y, we just use a basic RNN to present Z from the explicit causes.

To better illustrate our problem and objective, we provide a simple numerical example as follows:

Example 1. *We consider a temporal Boolean network with three Boolean variables: Y_t, X_t, and W_t, where Y_t at time t is determined by*

$$Y_t = X_{t-1} \text{ AND } W_{t-1}. \tag{11}$$

At every time step t, X_t and W_t are sampled from a uniform binomial distribution, i.e., $P(X_t = 0) = P(X_t = 1) = P(W_t = 0) = P(W_t = 1) = 0.5$.

In this context, the time-varying variables Y_t, X_{t-1}, and W_{t-1} at a particular time t are considered as the static random variables generated by the random process [1–3,22,24,28]. We list all the possible circumstances of the truth table of these three variables.

$$\begin{aligned} &P(X_{t-1}=0, W_{t-1}=0, Y_t=0)=1/4;\\ &P(X_{t-1}=0, W_{t-1}=1, Y_t=0)=1/4;\\ &P(X_{t-1}=1, W_{t-1}=0, Y_t=0)=1/4;\\ &P(X_{t-1}=1, W_{t-1}=1, Y_t=1)=1/4. \end{aligned} \tag{12}$$

Then, we calculate $H(Y_t)$, $I(Y_t; X_{t-1})$, and $I(Y_t; X_{t-1}, W_{t-1})$ respectively:

$$\begin{aligned} H(Y_t) &= -\sum_{y\in Y_t} p(y)\log(p(y)) = 2-\frac{3}{4}log3,\\ I(Y_t; X_{t-1}) &= \sum_{y\in Y_t}\sum_{x\in X_{t-1}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \frac{3}{2}-\frac{3}{4}\log 3,\\ I(Y_t; X_{t-1}, W_{t-1}) &= \sum_{y\in Y_t}\sum_{x\in X_{t-1}}\sum_{w\in W_{t-1}} p(w,x,y)\log\left(\frac{p(w,x,y)}{p(w,x)p(y)}\right) = 2-\frac{3}{4}\log 3. \end{aligned} \tag{13}$$

We notice that $H(Y_t) = I(Y_t; X_{t-1}, W_{t-1}) > I(Y_t; X_{t-1})$, which means that the information of Y_t could be completely covered when X_{t-1}, W_{t-1} are known. When W_{t-1} is unobserved, the observed X_{t-1} can only provide part of the information of Y_t. This gap of information is described by the conditional mutual information, $I(Y_t; W_{t-1}|X_{t-1})$, where

$$I(Y_t; W_{t-1}|X_{t-1}) = I(Y_t; X_{t-1}, W_{t-1}) - I(Y_t; X_{t-1}) = \frac{1}{2}, \tag{14}$$

which quantifies the information that W_{t-1} provides to Y_t that X_{t-1} cannot provide.

Example 2. *Now, we consider the scenario in which we do not know Y_t is generated by Equation (11) or the distribution of W_{t-1}, but we are provided with X_{t-1} and Y_t, and we wish to generate Z_{t-1} to complement $I(W_{t-1}; Y_t|X_{t-1})$ and maximize the objective in Equation (7). The ideal case is to reverse the Markov chain $(W_{t-1}, X_{t-1} \to Y_t)$ to $(X_{t-1}, Y_t \to W_{t-1})$ and make the posterior distribution $p(Z_{t-1}|X_{t-1}, Y_t)$ equal to $p(W_{t-1}|X_{t-1}, Y_t)$, where the joint and posterior distributions are*

$$\begin{aligned} &P(X_{t-1}=0, Y_t=0, Z_{t-1}=0)=1/4,\\ &P(X_{t-1}=0, Y_t=0, Z_{t-1}=1)=1/4,\\ &P(X_{t-1}=1, Y_t=0, Z_{t-1}=1)=1/4,\\ &P(X_{t-1}=1, Y_t=1, Z_{t-1}=0)=1/4, \end{aligned} \tag{15}$$

$$p(Z_{t-1}=1|X_{t-1}, Y_t) = p(W_{t-1}=1|X_{t-1}, Y_t) = \begin{cases} 1, & X_{t-1}=1, Y_t=1\\ 0, & X_{t-1}=1, Y_t=0\\ \frac{1}{2}, & X_{t-1}=0 \end{cases}. \tag{16}$$

In such a case, we calculate that

$$\begin{aligned} I(Y_t; X_{t-1}, Z_{t-1}) &= I(Y_t; X_{t-1}, W_{t-1}) = 2-\frac{3}{4}log3,\\ I(Y_t; Z_{t-1}|X_{t-1}) &= I(Y_t; W_{t-1}|X_{t-1}) = \frac{1}{2}. \end{aligned} \tag{17}$$

Example 3. *However, there actually exist more than one $p(Z_{t-1}|X_{t-1}, Y_t)$ that can maximize $I(Y_t; Z_{t-1}|X_{t-1})$. For example, we let*

$$Z_{t-1} = X_{t-1} \text{ XOR } Y_t, \tag{18}$$

and we can again list all the possible circumstances of the truth table of these three variables.

$$\begin{aligned} P(X_{t-1}=0, Y_t=0, Z_{t-1}=0) &= 1/2, \\ P(X_{t-1}=1, Y_t=0, Z_{t-1}=1) &= 1/4, \\ P(X_{t-1}=1, Y_t=1, Z_{t-1}=0) &= 1/4. \end{aligned} \tag{19}$$

In such a case, we can obtain the same result as that in Equation (17), which means a different Z from Equations (16) and (18) can achieve the optimal objective in (7). This indicates that they are providing the same conditional information that W provides to Y. The value of Y_t is known when $X_{t-1}=0$, but is uncertain when $X_{t-1}=1$. W provides the conditional information by helping to eliminate this uncertainty (or conditional entropy $H(Y|X)$), which can also be eliminated by not including W but including Z from Equation (16) or (18). Therefore, Z can express conditional information $I(W;Y|X)$.

4. Methods

4.1. Neural Causal Information Extractor

We name our architecture the Neural Causal Information Extractor (NCIE), which is composed of a generator and a discriminator. The architecture is shown in Figure 3.

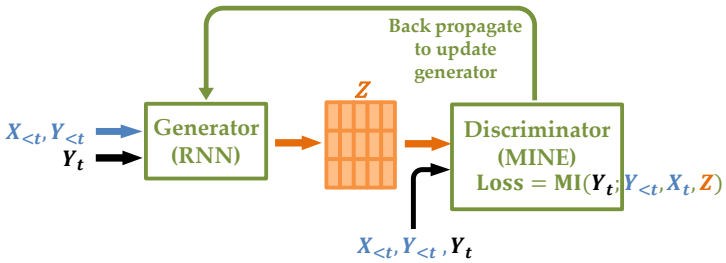

Figure 3. The architecture of the Neural Causal Information Extractor.

4.1.1. Generator

Although Z is not explicitly collected and observed from the system, we consider part of it to be involved in the system, specifically the part that provides information to Y. While we suppose that Y is generated by $\{X, Y_{<t}, Z\}$, reversely, the information of Z that we are interested in should be covered by (a part of) Y. Therefore, Z could be represented as a function of X, $Y_{<t}$, and Y. Other variables V besides $\{X, Y_{<t}, Y\}$ in the raw data should not be included to generate Z in case they have not provided information via path $V \to Z \to Y$ (or V should be detected as an observed direct cause before Z is included).

While finding the causes of Z is not the focus of this work (we only focus on finding the explicit and implicit causes of Y_t), we apply a basic RNN [18] to capture the lagged dependencies from X and Y to generate Z. The RNN consists of two neural networks to manipulate and propagate its hidden states. The first neural network, f_1, generates H_t from $X_{<t}$, $Y_{<t}$, Y_t, and H_{t-1} (Equation (20)), where H_t denotes the hidden states at time t. Then, the second neural network, f_2, reads H_t and generates the output Z (Equation (21)), while H_t is propagated to the next episode.

$$H_t = f_1(X_{t-1},\ Y_{t-1},\ H_{t-1}), \tag{20}$$

$$Z_t = f_2(H_t). \tag{21}$$

With the propagation of H_t, the memory of X and Y in previous episodes is also propagated. Furthermore, since Z is generated directly from H, the propagation of H also passes the memory and information of Z. It is worth noting here that Z is jointly generated by the explicit $\{X, Y\}$ and the implicit variables $Z_{<t}$. Suppose we employ another architecture that generates Z solely by explicit variables $\{X, Y\}$ (and without $Z_{<t}$);

Z then acts as an intermediate hidden layer in our architecture, which is equivalent to directly depicting Y by a neural network structure with $\{X, Y\}$ and without Z. In other words, Y would then be solely caused by explicit variables. This is why we apply RNN to generate Z jointly by implicits and explicits, which matches our assumption that Y is causally related to implicits and explicits.

4.1.2. Discriminator

The discriminator provides a measurement of $I(Y; X, Y_{<t}, Z)$, where Z is the output of the generator. Here, we consider Y and $\{X, Y_{<t}, Z\}$ as two sets of random variables, and their mutual information is equivalent to Kullback–Leibler (KL) divergence between their marginal and joint distributions,

$$\begin{aligned} I(Y; X, Y_{<t}, Z) &= D_{KL}(\mathbb{P}_{\boldsymbol{YXY_{<t}Z}} \,||\, \mathbb{P}_{\boldsymbol{Y}} \bigotimes \mathbb{P}_{\boldsymbol{XY_{<t}Z}}) \\ &= \sup_{T:\boldsymbol{Y}\times\boldsymbol{X}\times\boldsymbol{Y_{<t}}\times\boldsymbol{Z}\to\mathbb{R}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{YXY_{<t}Z}}}[T] - \log(\mathbb{E}_{\mathbb{P}_{\boldsymbol{Y}}\otimes\mathbb{P}_{\boldsymbol{XY_{<t}Z}}}[e^T])', \end{aligned} \tag{22}$$

where D_{KL} denotes the KL divergence of two distributions and $\mathbb{P}_{\boldsymbol{Y}}$ denotes the distribution of Y. The lower bound of KL divergence can be provided by Donsker–Varadhan representation (Equation (22)), where $T(Y, X, Y_{<t}, Z)$ is any class of function.

We follow MINE [39] to implement function T by a neural network with parameters Θ and maximize the lower bound of mutual information I_Θ by searching for optimized function T_θ:

$$I(Y; X, Y_{<t}, Z) \geq I_\Theta(Y; X, Y_{<t}, Z), \tag{23}$$

$$I_\theta(Y; X, Y_{<t}, Z) = \max_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_{\boldsymbol{YXY_{<t}Z}}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_{\boldsymbol{Y}}\otimes\mathbb{P}_{\boldsymbol{XY_{<t}Z}}}[e^{T_\theta}]). \tag{24}$$

Optimization is performed using stochastic gradient descent (SGD). However, a naive derivation of Equation (24) leads to biased gradient estimation. MINE [39] proposed addressing this bias by reformulating the loss function. For batch sample B, the gradient is

$$\widehat{G_B} = \mathbb{E}_B[\nabla_\theta T_\theta] - \frac{\mathbb{E}_B[\nabla_\theta T_\theta e^{T_\theta}]}{\mathbb{E}_B[e^{T_\theta}]}. \tag{25}$$

MINE suggests estimating the denominator of the second term using the moving average from previous epochs, while the other parts are the averaged gradient from the network parameters.

Among all (conditional) mutual information estimators, MINE estimates MI by stochastic gradient descent (some do not provide MI, and some are not trained by gradient descent). The available gradient is essential in our architecture because one of the inputs of MINE, Z, is exactly the output of the generator. Hence, the generator and discriminator are fused together by Z, which acts as the bridge to back-propagate the gradient from the discriminator to the generator and thus update the parameters in both neural networks.

4.1.3. NCIE: Maximizing Mutual Information

In Section 3, we claim that the objective of our architecture is to find implicit variables Z such that $I(Y; X, Y_{<t}, Z)$ is maximized. In this case, Z could represent the information provided by unobserved variables.

Since the generator and discriminator are fused together, they could be jointly considered as a whole neural network that trains both components simultaneously, where Z could be viewed as a layer in between that we are interested in. While training together by gradient descent, the generator is trained to provide better Z to push up I_θ, and the discriminator is trained to provide a better (higher) estimation of I_θ given the Z from the generator. Therefore, NCIE pursues objective

$$\max_{Z,\,\theta\,\in\,\Theta} I_\theta(Y; X, Y_{<t}, Z). \tag{26}$$

In an alternative view, the whole NCIE structure could be interpreted as a variant of an RNN that pursues to provide a maximized value of the Donsker–Varadhan representation given the batched inputs of X, $Y_{<t}$, and Y.

4.2. Verifying Causality from Z to Y by Time Series Prediction

4.2.1. Motivation and Architecture

To verify the capability of extracting effective implicit causes, we modify the NCIE in Section 4.1 and apply the architecture to perform multi-variate time series prediction in Section 5.2.2. To predict Y_t, we cannot use Y_t itself or any variables at time t or after t as the input; therefore, we have to find an alternative for the input of the generator. Here, we use a simple RNN to pre-train a $\hat{Y}_t$ that tries to estimate Y_t given all the explicit causes $\{X, Y_{<t}\}$ of Y_t. After the pre-training stage, $\hat{Y}_t$ is plugged into the input of the generator in the NCIE and replaces Y_t, to avoid the information leakage in the original architecture in Section 4.1.

The overall procedure is shown in Figure 4, which is summarized as follows:

1. We select the explicit causes X from the raw time series with a maximum lag of k using PCMCI [3], which is a framework that identifies the time-lagged causal relationship given the assumption of the completeness of the given data (no unobserved data). With its high precision, we assume the resulting X is the set of all the real causes of Y. Furthermore, we select the lagged $Y_{<t}$ from $t-k$ to $t-1$ as the autoregressive terms, together with X to be the set of observed causes;
2. We train the pre-training module and generate $\hat{Y}_t$ to estimate Y_t;
3. After finishing training the previous module, we use output $\hat{Y}_t$ as the input of this modified NCIE module and train the generator and discriminator (Section 4.1) together to generate Z;
4. We train neural network f_3 (Equation (27)) to generate $\hat{Y}_t{}'$, which is applied to examine the improvement of the prediction effect on Y given by Z.

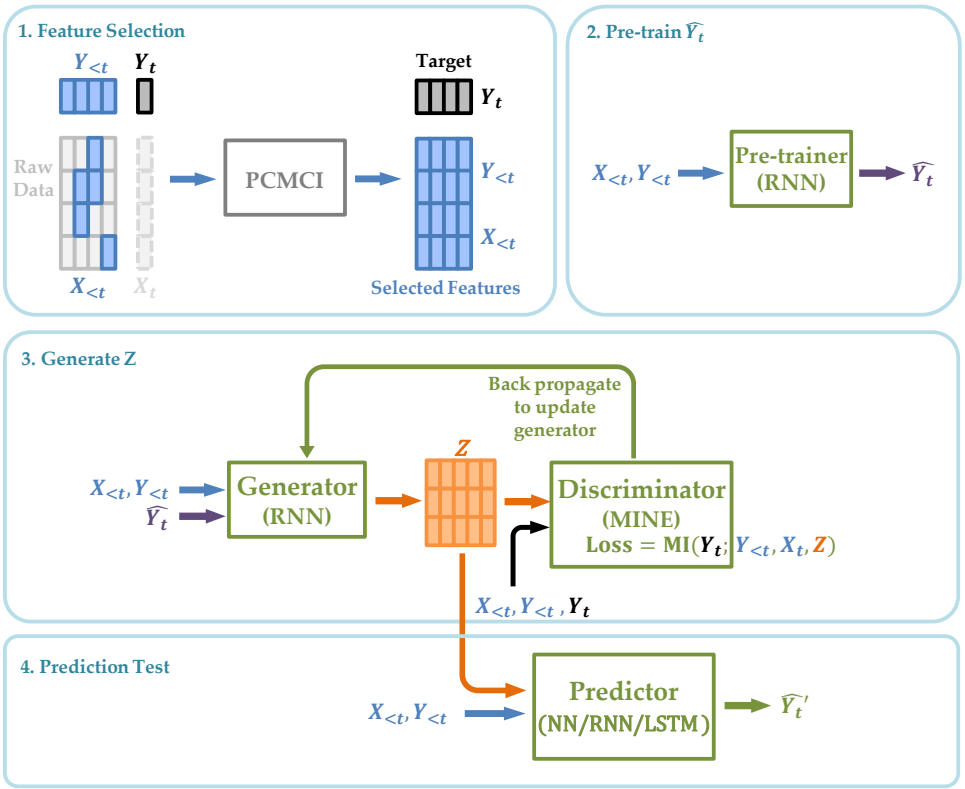

Figure 4. The architecture of Neural Causal Information Extractor to perform time series prediction.

We note that although there are four neural network architectures in Figure 4 and some of their inputs and outputs are connected, only the two networks in the third module train and update their parameters simultaneously. The framework in Figure 4 proceeds module by module, and the next one starts after the previous one finishes.

4.2.2. Prediction

We test whether the generated Z can help to improve the prediction of Y by comparing $\hat{Y}_t$ and $\hat{Y}_t{}'$. In Section 5.2.2, we employ different neural networks to implement f_3, which performs the prediction of Y by minimizing the mean squared error (MSE):

$$\hat{Y}_t{}' = f_3(X_t,\ Y_{t-1},\ Z_t). \tag{27}$$

5. Experiments

5.1. Synthetic Data Experiments: Dynamics of Z

In the following experiments on synthetic data, we demonstrate that the proposed NCIE framework can accurately extract the unobserved causal variables. We consider three synthetic systems (Figure 5) to compare the generated Z with the unobserved cause W that Z mimics. Targets Y in the synthetic systems are constructed with causes exhibiting long-term and short-term dynamics. We investigate the ability of NCIE in Section 4.1 to extract causes W from target Y when Y and the other causes X are given.

5.1.1. Case 1: A Linear System

We suppose the time-varying X and Y are observable while W is unobserved:

$$\begin{aligned} Y(t) &= X(t) + W(t), \\ X(t) &= \sin\frac{t}{10}, \\ W(t) &= \sin 10t. \end{aligned} \tag{28}$$

5.1.2. Case 2: A Non-Linear System

We suppose the time-varying X and Y are observable while W is unobserved:

$$\begin{aligned} Y(t) &= X(t) + W(t)(0.2 + X(t)), \\ X(t) &= 0.05t^3 - 15t^2 - 80t + 2, \\ W(t) &= \sin 10t. \end{aligned} \tag{29}$$

5.1.3. Case 3: A Non-Linear System

We suppose the time-varying X and Y are observable while W is unobserved:

$$\begin{aligned} Y(t) &= X(t) + 0.2W(t)(0.2 + X(t)), \\ X(t) &= 0.05t^3 - 15t^2 - 80t + 2, \\ W(t) &= \sin 10t. \end{aligned} \tag{30}$$

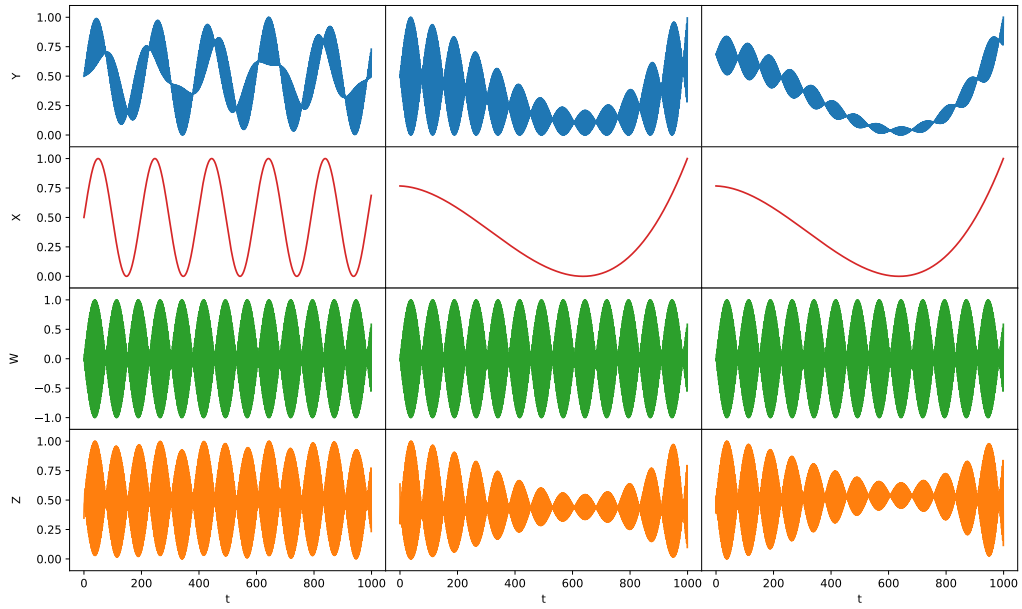

Figure 5. Dynamics of the normalized Y, X, W, and Z for Synthetic Case 1 (**Left**), Case 2 (**Middle**), Case 3 (**Right**).

5.1.4. Case 4: 1—Predator 2—Prey Model Where Prey Share the Same Food

We investigate our approach with the predator–prey model, which is a classical problem in ecology [43,44]. Specifically, we choose the model with 1 predator species, wolves, and 2 prey species, sheep and rabbits. The prey species share the same food source: grass. In total, there are four species in the ecosystem, and their populations grow according to the Lotka–Volterra model [45,46]:

$$\begin{aligned} \frac{dW(t)}{dt} &= W(t)(-a_0 + a_1 S(t) + a_2 R(t)), \\ \frac{dS(t)}{dt} &= S(t)(b_0 - b_1 W(t) + b_2 G(t)), \\ \frac{dR(t)}{dt} &= R(t)(c_0 - c_1 W(t) + c_2 G(t)), \\ \frac{dG(t)}{dt} &= G(t)(d_0 - d_1 S(t) - d_2 R(t)), \end{aligned} \tag{31}$$

where $W(t)$, $S(t)$, $R(t)$, and $G(t)$ represent the population of wolves, sheep, rabbits, and grass at time t, respectively. Because this dynamic system can lead to negative populations and we want to prevent extinction from halting our simulation, we add a modification that sets any population to 1 if it drops below 1.

We consider the scenario in which some intermediate species in the biological chain are unknown or unobserved. Therefore, we assume that $S(t)$ and $R(t)$ are unobserved. We set $W(t)$ as the target variable and $G(t-1)$ as the observed cause, and attempt to recover the information provided by the unobserved causes $S(t)$ and $R(t)$. It is worth noting that $G(t-1)$ is not a direct cause of $W(t)$ in this case. The absence of $S(t-1)$ and $R(t-1)$ makes $G(t-1)$ and $W(t)$ conditionally dependent, thus creating a spurious causal link. Since it is not the focus of our work, we do not discuss how to eliminate spurious links after recovering unobserved causes in this study.

The dynamics of the population of the four species and Z are shown in Figure 6. From this, we can observe that Z provides information about the unobserved variables $S(t)$ and $R(t)$.

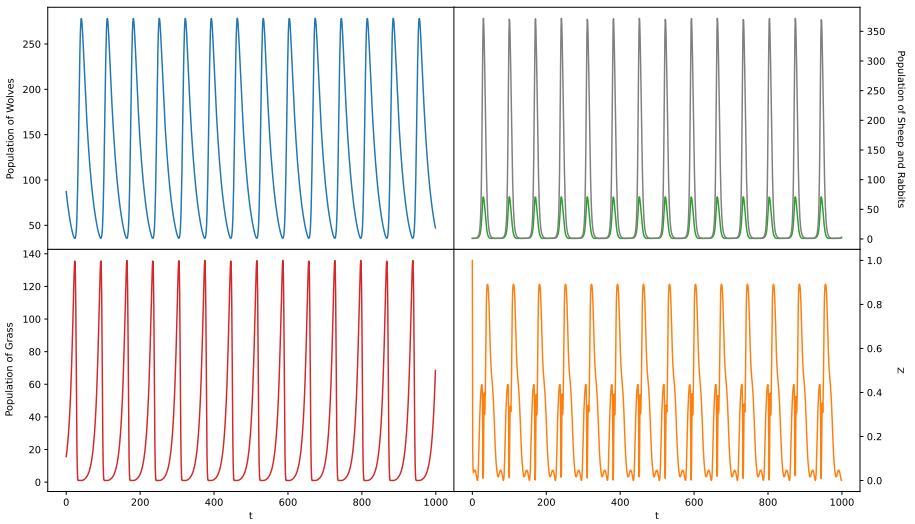

Figure 6. Dynamics of Z and the population of four species for Synthetic Case 4.

5.1.5. The Interpretation of Z

Our target is to generate Z that carries the unique information that unobserved causes provide to Y. However, in the three synthetic examples, instead of revealing the dynamic of W, Z exhibits a variable that tries to introduce Y while eliminating the information of X.

We try to explain the result using partial information decomposition theory [47], which suggests that we could decompose the Venn Diagram (Figure 2) of mutual information into the following parts shown in Figure 7:

$$\begin{aligned} H(Y) &= I(Y; X, Y_{<t}, W) \\ &= I_{\text{uni}}(Y; X, Y_{<t}) + I_{\text{uni}}(Y; W) \\ &\quad + I_{\text{red}}(Y; X, Y_{<t}, W) + I_{\text{syn}}(Y; X, Y_{<t}, W). \end{aligned} \tag{32}$$

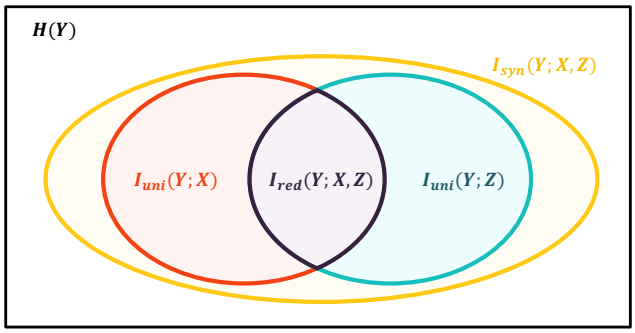

Figure 7. The Venn Diagram illustrates the partial information decomposition of Y into four discrete parts: the unique information from X (red), the unique information from Z (blue), the redundant information from X and Z (purple), and the synergistic information from X and Z (yellow).

The information to Y from all of its causes consists of four parts:

- $I_{\text{uni}}(Y; W)$, the unique information that could only be provided by W;
- $I_{\text{uni}}(Y; X, Y_{<t})$, the unique information that could only be provided by $\{X, Y_{<t}\}$;
- $I_{\text{syn}}(Y; X, Y_{<t}, W)$, the synergistic information provided only when both $\{X, Y_{<t}\}$ and W are present;

- $I_{\text{red}}(Y; X, Y_{<t}, W)$, the redundant information that could be provided by either $\{X, Y_{<t}\}$ or W.

If our objective is to recover W, Z should only contain the unique information $I_{\text{uni}}(Y; W)$ and the synergistic information $I_{\text{syn}}(Y; X, Y_{<t}, W)$, while the unique information $I_{\text{uni}}(Y; X, Y_{<t})$ and the redundant information $I_{\text{red}}(Y; X, Y_{<t}, W)$ are not present in this case. However, our objective in generating Z is not to exactly mimic W, but to maximize the information that the observed and unobserved causes provide to Y, so that we can acquire a better causal representation for Y.

Moreover, it is difficult to integrate different methods; therefore, our architecture does not measure the four components [48–50] separately but rather measures and maximizes the entire $I(Y; X, Y_{<t}, Z)$ at once. Thus, $I_{\text{uni}}(Y; Z)$, $I_{\text{syn}}(Y; X, Y_{<t}, Z)$, and $I_{\text{red}}(Y; X, Y_{<t}, Z)$ in Equation (32) are maximized. As a result, the implicit causes Z may contain all four components, which is the information from both the observed $\{X, Y_{<t}\}$ and the unobserved W.

Therefore, in the synthetic cases, we still observe the trend information from X that is not fully eliminated in Z. Still, in this case, we argue that Z carries significant information and reveals evident dynamics of W. As shown in Table 1, the observed causes X only provide very little information to the target Y in simple synthetic cases, where the mutual information of target Y provided by itself, $I(Y; Y)$, is equivalent to the entropy of Y, $H(Y)$, provided that we consider Y as a random variable. In contrast, after we add the implicit causes Z, all the causes together provide most of the information of target Y and thus complete the causal structure of Y.

Table 1. Mutual information provided to Y from different sets of variables.

Case	Explicit X	Explicit X + Implicit Z	Target Y
1	0.646	5.46	5.52
2	0.679	4.11	5.35
3	1.44	3.47	5.29
4	2.84	3.48	5.36

5.2. Preformance Evaluation by Real-World Data Experiments

5.2.1. Information Recovery

We first examine the ability of NCIE (Section 4.1) to extract implicit causes that recover unobserved information of the targets in several real-world datasets, including

- The Electricity Transformer (Oil) Temperature (ETT) [17], which includes four datasets sampled at different time intervals (1 min, 2 min, 1 h, and 2 h). The observed causes include useful and useless loads in different levels;
- The daily exchange rates [51] of eight countries, with one of them set as the target and the other seven as the observed causes for predicting the target;
- Minneapolis–St Paul interstate metro traffic volume [52], in which observed causes include temperature, weather, and holidays;
- The PM2.5 quantity in Beijing (https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate) (accessed on 6 November 2023), in which observed causes include dew, temperature, atmospheric pressure, weather, as well as wind direction and speed (all the codes and data are available at https://github.com/jh-liang/NeuralCausalInfoExtractor) (accessed on 6 November 2023).

These data provide partial observations of complete systems in industrial monitoring, financial markets, and urban operations. In the provided datasets, it is obvious that some causes, W, are unobserved. For example, exchange rates are influenced not only by foreign exchange rates but also by the domestic economy and complex international trade. Similarly, the PM2.5 datasets do not involve urban human activities and pollution emissions from power plants. We aim to recover conditional information $I(W; Y|X)$ using Z, which advances the precision of time series forecasting.

Table 2 shows the information provided by only observed causes and by the combination of observed and implicit causes. It illustrates that the generated Z can provide significant extra information besides the observed causes X in real-world scenarios. The Z in this table is generated by NCIE (Section 4.1). We do not discuss the dimensionality of Z in this paper. While the higher dimensionality of Z provides the wider information channel, we simply set the dimension of Z to be the dimension of $\{X, Y_{<t}\}$.

Table 2. Mutual information provided to Y from different sets of variables. The Z in this table is generated by the original NCIE (Section 4.1).

Datasets	Explicit	Explicit + Implicit Z	Target Y
ETTh1	2.78	5.54	6.75
ETTh2	3.55	5.55	6.66
ETTm1	3.12	5.46	6.33
ETTm2	4.278	5.36	6.27
ExcRat	3.63	6.24	7.31
Metro	2.66	5.68	7.37
PM2.5	1.80	4.59	5.52

5.2.2. Single-Step Time Series Forecasting

While we claim that our architecture helps to identify unobserved causes of the target variables, it holds massive potential for extracting useful information from the available features and for conducting time series forecasting. We evaluate our architecture by performing time series forecasting on the real-time series datasets and comparing our results with the baselines [18,19] listed in Table 3. To prevent future information leakage in time series prediction, we employ the modified NCIE architecture in Section 4.2 to generate Z.

In this work, we are considering one-dimensional targets Y with multiple features that are candidates for observed causes chosen by PCMCI. Additionally, our architecture focuses on single-step time series forecasting because our goal is to find the direct causes for Y, which could be X_{t-k} and Z_{t-k} with a small lag k. The results are shown in Table 3. For the three chosen time series analyzing tools, we compare the prediction results with and without Z. It is illustrated that including Z in most cases does outperform those predictions without implicit variables Z.

Table 3. Single-step time series forecasting with and without Z in three different architectures, where 'NN' denotes the result from a simple NN with input including only observed causes, and 'I-NN' denotes the result with input including both observed causes and implicit variables Z generated by modified NCIE. The better result in the comparison is underlined.

Datasets	NN	I-NN	RNN	I-RNN	LSTM	I-LSTM
ETTh1	1.97×10^{-4}	$\underline{1.89 \times 10^{-4}}$	$\underline{1.65 \times 10^{-4}}$	1.77×10^{-4}	2.05×10^{-4}	$\underline{2.01 \times 10^{-4}}$
ETTh2	1.81×10^{-4}	$\underline{1.45 \times 10^{-4}}$	2.05×10^{-4}	$\underline{1.28 \times 10^{-4}}$	2.62×10^{-4}	$\underline{1.64 \times 10^{-4}}$
ETTm1	4.19×10^{-5}	$\underline{4.16 \times 10^{-5}}$	4.35×10^{-5}	$\underline{4.24 \times 10^{-5}}$	7.57×10^{-5}	$\underline{5.12 \times 10^{-5}}$
ETTm2	2.22×10^{-5}	$\underline{1.68 \times 10^{-5}}$	4.59×10^{-5}	$\underline{3.37 \times 10^{-5}}$	1.45×10^{-4}	$\underline{8.17 \times 10^{-5}}$
ExcRat	9.90×10^{-5}	$\underline{9.84 \times 10^{-5}}$	2.01×10^{-4}	$\underline{1.99 \times 10^{-4}}$	4.25×10^{-4}	$\underline{3.57 \times 10^{-4}}$
Metro	3.33×10^{-3}	$\underline{2.91 \times 10^{-3}}$	4.09×10^{-3}	$\underline{4.03 \times 10^{-3}}$	4.36×10^{-3}	$\underline{3.94 \times 10^{-3}}$
PM2.5	5.50×10^{-4}	$\underline{5.46 \times 10^{-4}}$	5.87×10^{-4}	$\underline{5.78 \times 10^{-4}}$	5.87×10^{-4}	$\underline{5.70 \times 10^{-4}}$

Our problem could be considered a feature selection task. The Z we generate could be considered as a feature that helps to describe Y. We see that in a few cases, Z does not necessarily improve the prediction of Y. This could be due to two reasons. First, although we know there is a gap between the entropy of Y and the information that X and $Y_{<t}$ bring to Y, this gap might be too narrow for NCIE to generate a Z that brings sufficient information from it. Second, when the system is time-varying, the underlying causality

structure of the system could also vary with time. Z that acts as a significant cause in (part of) the training dataset may become a redundant variable that perturbs the prediction.

It is worth mentioning that RNN and LSTM do not outperform simple neural networks (NN) in all cases. Our explanation is that the features X that PCMCI chooses cover most of the (observed) direct causes of Y, and we also choose autoregressive features $Y_{<t}$ from $t-k$ to $t-1$ that already cover short- and long-term memories of Y. Therefore, the memories that RNN and LSTM extract could not significantly help the prediction, but, in contrast, overfit Y.

Table 4 shows the mutual information between different sets of variables and Y. For real data, the Z that NCIE generates can provide extra information to Y. It demonstrates NCIE's ability to extract part of the implicit causes in the system, although we find a gap between $I(Y; X, Y_{<t}, Z)$ and $I(Y; Y)$. It is unavoidable in cases in which the systems are open or noisy because external noises are not driven by variables within the system. Furthermore, for a large system with too many causes, some causes may carry very little unique information to the target, and it is difficult to express them with limited Z.

Table 4. Mutual information provided to Y from different sets of variables. The Z in this table is generated by the modified NCIE (Section 4.2).

Datasets	Pre-Trained $\hat{Y}$	Explicit	Explicit + Z	Y
ETTh1	1.46	3.59	3.60	6.31
ETTh2	2.56	3.65	3.90	6.69
ETTm1	2.51	3.05	2.94	5.58
ETTm2	3.39	4.36	4.37	6.23
ExcRat	2.87	2.98	4.14	6.88
Metro	1.06	3.42	3.45	7.60
PM2.5	0.96	2.32	2.34	6.53

The first column of Table 4 shows the mutual information between the pre-trained $\hat{Y}$ and the actual Y. We recall that the pre-trained $\hat{Y}$ is generated by an RNN architecture that minimizes the mean square error (MSE), given the explicit causes X and $Y_{<t}$. The gap from the first column to the others shows that RNN and other recurrent structures do not necessarily provide a regression result with sufficient information to the target. While it is not a key point to argue the effectiveness of different objectives in this work, it reminds us that NCIE provides more information than general neural architectures.

6. Conclusions

In this work, we propose NCIE to generate implicit variables that complement the unobserved information of the target not provided by the observed variables. We can complete the causal structure of the target with the implicit variables while retaining the observed causes, as they together provide most of the target's information. Furthermore, the generated implicit variables have similar dynamics to the unobserved causes in the synthetic experiments and help to predict the target in the real-world time series. These results provide sufficient evidence that the implicit variables can be an effective candidate to substitute the unobserved causes and complete the causal structure.

While this work only focuses on discovering unobserved causes for single targets, future attempts could be made to apply similar methods to construct a complete graph for multiple targets, in which all the causes provide complete information for all the targets. With such methods, we are able to depict the evolution of systems by the information flow among variables, which could contribute to the research of emergence in real-world complex systems.

Author Contributions: Conceptualization, K.-H.L. and Y.X.; Data curation, K.-H.L.; Formal analysis, K.-H.L.; Funding acquisition, B.C. and W.K.C.; Investigation, K.-H.L.; Methodology, K.-H.L. and Y.X.; Project administration, B.C. and W.K.C.; Resources, B.C. and W.K.C.; Software, K.-H.L. and Y.X.;

Supervision, B.C. and W.K.C.; Validation, K.-H.L., Y.X., B.C. and W.K.C.; Visualization, K.-H.L. and Y.X.; Writing—original draft, K.-H.L.; Writing—review and editing, K.-H.L., Y.X., B.C. and W.K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Innovation Commission of Shenzhen (JCYJ20210324135011030, JCYJ20210324115604012, WDZC20200818121348001), the National Natural Science Foundation of China (71971127), the Guangdong Pearl River Plan (2019QN01X890), the High-End Foreign Expert Talent Introduction Plan (G2021032022L), the Tsinghua Shenzhen International Graduate School Fund (JC2021004), and the Science and Technology Innovation Committee of Shenzhen-Platform and Carrier (International Science and Technology Information Center).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All the codes and data are available at https://github.com/jh-liang/NeuralCausalInfoExtractor (accessed on 6 November 2023)

Acknowledgments: We gratefully acknowledge Keqin Guan, Kexin Cao, Siyu Chen, and Chang-Yan Shih for their tremendous advice on ideas and academic writing. Likewise, we thank the support from the Causal Emergence Reading Group supported by the Save 2050 Programme jointly sponsored by Swarma Club and X-Order.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Gong, C.; Yao, D.; Zhang, C.; Li, W.; Bi, J. Causal discovery from temporal Data: An overview and new perspectives. *arXiv* **2023**, arXiv:2303.10112.
2. Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, Prediction, and Search*; MIT Press: Cambridge, MA, USA, 2000.
3. Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **2019**, *5*, eaau4996. [CrossRef]
4. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.* **1995**, *20*, 197–243. [CrossRef]
5. Kayaalp, M.; Cooper, G.F. A Bayesian network scoring metric that is based on globally uniform parameter priors. *arXiv* **2012**, arXiv:1301.0576.
6. Marcinkevičs, R.; Vogt, J.E. Interpretable models for Granger causality using self-explaining neural networks. *arXiv* **2021**, arXiv:2101.07600.
7. Jiang, P.; Kumar, P. Information transfer from causal history in complex system dynamics. *Phys. Rev. E* **2019**, *99*, 012306. [CrossRef] [PubMed]
8. Li, J.; Convertino, M. Inferring ecosystem networks as information flows. *Sci. Rep.* **2021**, *11*, 7094. [CrossRef]
9. Engelberg, J.E.; Parsons, C.A. The causal impact of media in financial markets. *J. Financ.* **2011**, *66*, 67–97. [CrossRef]
10. Farag, H.; Cressy, R. Do unobservable factors explain the disposition effect in emerging stock markets? *Appl. Financ. Econ.* **2010**, *20*, 1173–1183. [CrossRef]
11. Williams, B.K.; Brown, E.D. Partial observability and management of ecological systems. *Ecol. Evol.* **2022**, *12*, e9197. [CrossRef]
12. Chadès, I.; Pascal, L.V.; Nicol, S.; Fletcher, C.S.; Ferrer-Mestres, J. A primer on partially observable Markov decision processes (POMDPs). *Methods Ecol. Evol.* **2021**, *12*, 2058–2072. [CrossRef]
13. Singh, M.F.; Wang, A.; Braver, T.S.; Ching, S. Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling. *J. Neural Eng.* **2020**, *17*, 046025. [CrossRef] [PubMed]
14. Gupta, V.; Li, L.K.; Chen, S.; Wan, M. Model-free forecasting of partially observable spatiotemporally chaotic systems. *Neural Netw.* **2023**, *160*, 297–305. [CrossRef]
15. Duan, C.; Jiang, Y.; Pu, H.; Luo, J.; Liu, F.; Tang, B. Health prediction of partially observable failing systems under varying environments. *ISA Trans.* **2023**, *137*, 379–392. [CrossRef] [PubMed]
16. Geiger, P.; Zhang, K.; Schoelkopf, B.; Gong, M.; Janzing, D. Causal inference by identification of vector autoregressive processes with hidden components. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1917–1925.
17. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.
18. Lipton, Z.C.; Berkowitz, J.; Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv* **2015**, arXiv:1506.00019.
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

20. Yao, W.; Sun, Y.; Ho, A.; Sun, C.; Zhang, K. Learning temporally causal latent processes from general temporal data. *arXiv* **2021**, arXiv:2110.05428.
21. Klindt, D.; Schott, L.; Sharma, Y.; Ustyuzhaninov, I.; Brendel, W.; Bethge, M.; Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv* **2020**, arXiv:2007.10930.
22. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2018**, arXiv:1808.06670.
23. Hyvärinen, A.; Shimizu, S.; Hoyer, P.O. Causal modelling combining instantaneous and lagged effects: An identifiable model based on non-Gaussianity. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 424–431.
24. Hyvarinen, A.; Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*; Singh, A., Zhu, J., Eds.; MIT Press: Cambridge, MA, USA, 2017; Volume 54, pp. 460–469.
25. Clark, D.; Livezey, J.; Bouchard, K. Unsupervised discovery of temporal structure in noisy data with dynamical components analysis. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
26. Bai, J.; Wang, W.; Zhou, Y.; Xiong, C. Representation learning for sequence data with deep autoencoding predictive components. *arXiv* **2020**, arXiv:2010.03135.
27. Meng, R.; Luo, T.; Bouchard, K. Compressed predictive information coding. *arXiv* **2022** arXiv:2203.02051.
28. Wu, H.; Gattami, A.; Flierl, M. Conditional mutual information-based contrastive loss for financial time series forecasting. In Proceedings of the First ACM International Conference on AI in Finance, New York, NY, USA, 15–16 October 2020; pp. 1–7.
29. Granger, C.W. Investigating causal relations by econometric models and cross-spectral methods. *Econom. J. Econom. Soc.* **1969**, *37*, 424–438. [CrossRef]
30. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5.
31. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
32. Rosas, F.E.; Mediano, P.A.; Jensen, H.J.; Seth, A.K.; Barrett, A.B.; Carhart-Harris, R.L.; Bor, D. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Comput. Biol.* **2020**, *16*, e1008289. [CrossRef] [PubMed]
33. 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; pp. 23–47.
34. Gerhardus, A.; Runge, J. High-recall causal discovery for autocorrelated time series with latent confounders. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12615–12625.
35. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
36. Louizos, C.; Shalit, U.; Mooij, J.M.; Sontag, D.; Zemel, R.; Welling, M. Causal effect inference with deep latent-variable models. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
37. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
38. Xiu, Y.; Cao, K.; Ren, X.; Chen, B.; Chan, W.K. Self-similar growth and synergistic link prediction in technology-convergence networks: The case of intelligent transportation systems. *Fractal Fract.* **2023**, *7*, 109. [CrossRef]
39. Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. Mine: Mutual information neural estimation. *arXiv* **2018**, arXiv:1801.04062.
40. Mukherjee, S.; Asnani, H.; Kannan, S. CCMI: Classifier based conditional mutual information estimation. In Proceedings of the Uncertainty in Artificial Intelligence, Virtual, 3–6 August 2020; pp. 1083–1093.
41. Zhang, R.; Koyama, M.; Ishiguro, K. Learning structured latent factors from dependent data: A generative model framework from information-theoretic perspective. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 11141–11152.
42. Zhu, H.; Wang, S. Learning fair models without sensitive attributes: A generative approach. *arXiv* **2022**, arXiv:2203.16413.
43. Diz-Pita, É.; Otero-Espinar, M.V. Predator–prey models: A review of some recent advances. *Mathematics* **2021**, *9*, 1783. [CrossRef]
44. Leeuwen, E.v.; Jansen, V.; Bright, P. How population dynamics shape the functional response in a one-predator–two-prey system. *Ecology* **2007**, *88*, 1571–1581. [CrossRef] [PubMed]
45. Lotka, A.J. *Elements of Physical Biology*; Williams & Wilkins: Philadelphia, PA, USA, 1925.
46. Volterra, V. Volume 2, Societá anonima tipografica "Leonardo da Vinci". In *Variazioni e Fluttuazioni del Numero d'Individui in Specie Animali Conviventi*; Accademia Nazionale dei Lincei: Roma, Italy, 1927.
47. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
48. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
49. Kleinman, M.; Achille, A.; Soatto, S.; Kao, J.C. Redundant information neural estimation. *Entropy* **2021**, *23*, 922. [CrossRef]
50. Quax, R.; Har-Shemesh, O.; Sloot, P.M. Quantifying synergistic information using intermediate stochastic variables. *Entropy* **2017**, *19*, 85. [CrossRef]

51. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 95–104.
52. Hogue, J. Metro Interstate Traffic Volume. *Uci Mach. Learn. Repos.* **2019**. (accessed on 21 December 2023). [CrossRef]