Neural-HATS: Neural Hybrid Approach for Time Series Causal Discovery

Saima Absar[1], **Wen Huang**[1], **Yongkai Wu**[2], **Lu Zhang**[1]

[1]University of Arkansas, [2]Clemson University
sa059@uark.edu, wenhuang@uark.edu, yongkaw@clemson.edu, lz006@uark.edu

Abstract

This paper presents Neural-HATS (Neural Hybrid Approach for Time Series Causal Discovery), an innovative framework that combines conditional independence (CI) testing with continuous optimization-based learning algorithms to uncover causal structures in time series data. The approach features an attention-based encoder-decoder architecture integrated with Kernel Conditional Independence (KCI) testing, enabling direct CI tests between time series. These tests are then integrated into continuous optimization learning algorithms for enhanced causal discovery. The integration not only refines the CI process but also expands the capabilities of continuous optimization algorithms. Our experiments demonstrate the efficacy of this hybrid approach in deriving more accurate causal graphs, showing promise for extensive applications across various domains where time series data is prevalent.

Introduction

Causal discovery algorithms identify cause-and-effect relationships within data, represented as causal graphs. These algorithms are generally classified into two types: constraint-based and score-based methods (Guo et al. 2020; Hasan, Hossain, and Gani 2023; Zanga, Ozkirimli, and Stella 2022). Constraint-based methods use conditional independence tests to define causal directions under a fixed set of constraints (Krich et al. 2020). This approach often has high computational costs due to combinatorial optimization. Score-based methods, however, learn causal graphs by optimizing a score function that evaluates possible structures (Absar, Wu, and Zhang 2023; Marcinkevičs and Vogt 2021; Pamfil et al. 2020; Sun et al. 2021). Recent score-based techniques utilize continuous optimization and deep learning for cost function minimization (Vowels, Camgoz, and Bowden 2022), but they require large datasets, which are often challenging to obtain in practice.

The majority of causal discovery literature primarily addresses non-temporal, static data, yet real-world data often take the form of time series, where observations occur over time. Handling temporal data to uncover causal relationships is crucial in advancing causal discovery techniques (Assaad, Devijver, and Gaussier 2022b; Glymour, Zhang, and Spirtes 2019; Hasan, Hossain, and Gani 2023; Moraffah et al. 2021). A key approach for time series causal inference is Granger causality, which states that one time series can causally influence another if past values of the former predict future values of the latter, assuming causes precede effects and no hidden confounders exist (Granger 1969, 1980, 2001; Shojaie and Fox 2022). Early Granger causality methods used joint probability (Hiemstra and Jones 1994) and linear regression models (Luo et al. 2015), evolving later to non-linear forms through neural networks, such as in TCDF (Nauta, Bucur, and Seifert 2019), which uses attention-based convolutional neural networks (CNNs) to predict causality. Score-based approaches applying Granger causality include methods like DYNOTEARS, GVAR, NTS-NOTEARS, NTiCD (Absar, Wu, and Zhang 2023; Marcinkevičs and Vogt 2021; Pamfil et al. 2020; Sun et al. 2021), etc.

Despite advancements in time series causal discovery, constraint-based approaches grounded in graphical criteria are relatively scarce. This is likely due to the difficulties in performing Conditional Independence (CI) tests for time series data. Traditional CI tests, such as Pearson's chi-square, Fisher's exact test, or kernel-based methods like the Kernel Conditional Independence (KCI) test (for high-dimensional settings) (Zhang et al. 2012) are not applicable to temporal data. Methods such as PCMCI (Runge 2018) adapt CI testing to time series by treating each time point as a variable and testing for independence across time points, conditioned on other points. Yet, such methods often struggle with data scarcity at each time point or depend on strong assumptions about causal relationships, including stationarity.

To bridge this gap, we introduce a novel CI testing framework tailored for time series analysis. The core idea is to encode the predictive information contained in the time series into vector representations, and then leverage kernel-based approaches like the KCI to test for conditional independence directly between different time series. To compute the vector representations, we propose an attention-based encoder-decoder architecture that utilizes long-short-term memory (LSTM) networks as the encoder and decoder and learns hidden representations from input time series based on the attention mechanism. Specifically, to conduct the CI test $CI(\alpha, \beta \mid C)$ where α, β are time series and C is a set of

conditioned time series, we feed α and C into the encoding LSTM, and β into the decoding LSTM, to compute hidden features. These learned features are used to calculate attention scores, and the scored features then are aggregated and fed into a multilayer perceptron (MLP) for decoding, predicting the output time series β. Following training, the encoder and decoder are directly used to generate vector representations of the time series data, which then undergo a KCI test to assess conditional independence.

Similar to other constraint-based methods, a naive implementation that exhaustively tests conditional independence for all possible time series combinations incurs exponential time complexity. In this work, we show that our CI testing method can be innovatively integrated into the continuous optimization framework, resulting in a hybrid algorithm. Our approach employs low-order CI tests to derive a CI matrix, which can be integrated as a regularization constraint in the loss function of any score-based continuous optimization method that discovers a causal structure from time series data. This approach leverages CI tests to guide a continuous optimization process, while efficiently bypassing the need for exhaustive testing. Our experiments demonstrate that 1-order or even 0-order CI tests can effectively improve the performance of state-of-the-art continuous optimization algorithms. As a hybrid approach, our method offers several advantages. First, leveraging the strengths of deep neural networks and the KCI test, our proposed CI test method avoids assumptions about specific lag structures, stationarity, or linearity in the data. Second, by combining CI tests with continuous optimization, our method harnesses the efficiency of continuous optimization for causal structure search. Finally, our framework has the potential to further benefit from advancements in both CI testing and continuous optimization algorithms.

Preliminaries

Kernel-based Conditional Independence Test. In this paper, we utilize the Kernel-based Conditional Independence (KCI) test for conducting conditional independence testing. Conditional independence tests can be challenging when applied to continuous datasets, due to issues such as the curse of dimensionality or unknown data distributions. To address these challenges, a computationally efficient method has been proposed in (Zhang et al. 2012). This method leverages kernel matrices of the variables to define a simple test statistic. The authors demonstrated that independence and conditional independence can be characterized by the uncorrelatedness between functions in certain kernel spaces. They proposed a procedure involving the calculation of centralized kernel matrices and their eigenvalues for continuous variables. The test statistic is then evaluated from the trace of the kernel matrices. In our method, we used the RBF functions for kernel computations and implemented the Monte Carlo KCI tests with a significance level of 1%. The detail is provided in Appendix B.

Assumptions. In our method, we adopt the following assumptions that are commonly utilized in constraint-based causal discovery.

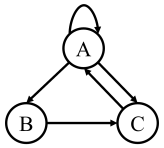


Figure 1: Example of a summary causal graph consisting of both a self-loop and cycles.

Assumption 1 (Causal Markov)**.** *Given a set of time series $\mathcal{X}$, and a directed graph $\mathcal{G}$, if there exist no direct edge from α to β in $\mathcal{G}$, then β is conditionally independent of α given $\mathbf{C}\backslash\{\alpha\}$ in $\mathcal{X}$.*

Assumption 2 (Causal Faithfulness)**.** *Given $\mathcal{X}$ and $\mathcal{G}$, if β is conditionally independent of α given $\mathbf{C}\backslash\{\alpha\}$ in $\mathcal{X}$, then there exists no direct edge from α to β in $\mathcal{G}$.*

The above two assumptions require that $\mathbf{C}$ does not contain any descendants (causally downstream variables) of β. Causal Markov implies that any given point in a time series is conditionally independent of all other points, given the past values of its parents (direct causes) for any lag. On the other hand, Causal Faithfulness implies that the only independencies that exist are those that can be explained by the temporal causal structure. That is, if two points in time are statistically independent, it is because there is no direct causal influence between them for any lag in the time series model. Based on these assumptions, constraint-based approaches establish a connection between the existence of direct edges and the level of conditional independence to identify the causal structure from data.

Methodology

Problem Statement

Consider $\mathcal{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ comprising d time series, each with a consistent length of n. Each time series $X^{(i)}$ is represented as a sequence $X^{(i)} = (x_0^{(i)}, x_1^{(i)}, \ldots, x_{n-1}^{(i)})$. A causal graph for time series, denoted as $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, can be characterized by a directed graph, where $\mathbf{V}$ represents the set of nodes, and $\mathbf{E}$ denotes the set of directed edges. Each node $X^{(i)} \in \mathbf{V}$ corresponds to a specific time series in $\mathcal{X}$, and each edge $(X^{(i)}, X^{(j)}) \in \mathbf{E}$ signifies a direct causal relationship from $X^{(i)}$ to $X^{(j)}$. There are various types of causal graphs for time series, depending on the level of granularity of the temporal dependencies represented in the graph. One commonly used one is the summary causal graph.

Definition 1 (Summary Causal Graph)**.** *Given multivariate time series $\mathcal{X}$, a summary causal graph is a directed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where:*

$$\mathbf{V} \coloneqq \{X^{(i)} \mid 1 \leq i \leq d\}$$

$$\mathbf{E} \coloneqq \{X^{(i)} \to X^{(j)} \mid \textit{if any past values of } X^{(i)} \textit{ causally influence } X^{(j)} \textit{ with any lag}\}$$

Note that, unlike a directed acyclic graph, a summary causal graph may have self-loops (i.e., a directed edge that starts

and ends at the same node) and cycles (i.e., a directed path that starts and ends at the same node).

A summary causal graph is often represented by an adjacency matrix $\mathcal{A} \in \mathbb{R}^{d\times d}$, where $\mathcal{A}_{i,j} = 1$ denotes an edge from $X^{(i)}$ to $X^{(j)}$. An illustrative example of a summary causal graph is shown in Fig. 1, where A, B, and C represent three time series variables. In this example, A is caused by C and itself, B is caused by A alone and C is caused by both A and B. Note that this graph features a self-loop at A and cycles such as $A \to B \to C \to A$, $A \to C \to A$, etc.

In this paper, we aim to uncover causal relationships in multivariate time series $\mathcal{X}$ and represent them in terms of a summary causal graph. Our hybrid approach involves three main steps:

1. Encode hidden information from the time series into vector representations using an attention-based encoder-decoder framework.
2. Utilize the KCI to test for conditional independence among variables and form a conditional independence matrix based on the results.
3. Integrate this matrix into a continuous optimization causal discovery method as a regularization term, to derive the summary causal graph.

An overview of our proposed approach is outlined in Fig. 2. The details of each of these steps will be discussed in the subsequent sections.

Conditional Independence Test

Constraint-based approaches rely on examining conditional independence relations among the variables of multivariate time series data. For time series, conditional independence between two time series α, β given a subset of time series $\mathbf{C}$ is defined as follows (Mogensen, Hansen et al. 2020).

Definition 2 (Conditional independence)**.** *Let $\alpha, \beta \in \mathbf{V}$, $\mathbf{C} \subseteq \mathbf{V}\backslash\{\alpha\}$. We say that β is conditionally independent of α given $\mathbf{C}$ if for any time point t, the past of $\mathbf{C}$ until time t gives us the same predictable information about β_{w_0} as the past of both α and $\mathbf{C}$ until time t, denoted by $\alpha \not\to \beta \mid \mathbf{C}$.*

We propose a novel conditional independence (CI) testing framework for time series by first encoding the time series into vector embeddings and then utilizing the KCI to test for conditional independence relationships, as detailed in the next subsection. To establish a theoretical basis for our method, we leverage the conditional transfer entropy $CTE_{\alpha\to\beta|\mathbf{C}}$ to quantify the amount of information flow from α to β specifically due to α and not due to $\mathbf{C}$, or that cannot be otherwise explained by the past values of β and $\mathbf{C}$ alone. We show that conditional independence tested based on the above embeddings implies that the corresponding conditional transfer entropy equals 0, which further implies conditional independence between time series.

Specifically, the conditional transfer entropy (CTE) could be expressed as:

$$CTE_{\alpha\to\beta|\mathbf{C}} = \mathcal{H}(\beta_{w_0}|\mathbf{C}_{w_1}) - \mathcal{H}(\beta_{w_0}|\alpha_{w_1}, \mathbf{C}_{w_1}) \quad (1)$$

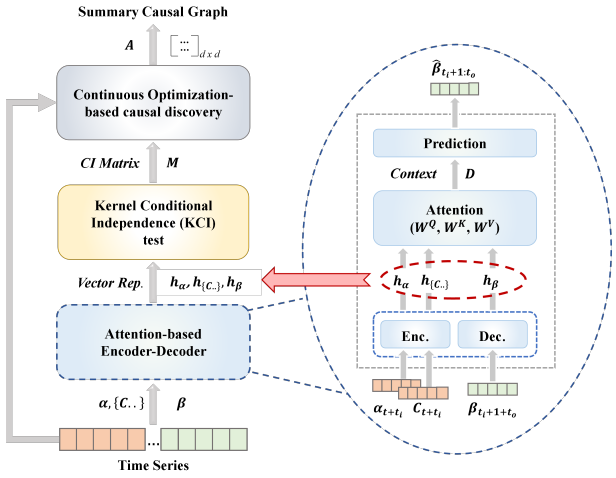


Figure 2: Overview of proposed hybrid temporal causal discovery, comprising three main components: an attention-based encoder-decoder, kernel-based conditional independence testing, and continuous optimization-based causal discovery method, to derive a summary causal graph A, with d denoting the number of variables. The architecture of the encoder-decoder model is depicted on the right.

where β_{w_0} is the time series in a time window from time t with length w_0, $\alpha_{w_1}, \mathbf{C}_{w_1}$ are the time series in a time window prior to time t with length w_1, and $\mathcal{H}$ is the conditional entropy given by

$$\mathcal{H}(\beta_{w_0}|\mathbf{C}_{w_1}) = -\sum_{\beta_{w_0},\mathbf{C}_{w_1}} P(\beta_{w_0}, \mathbf{C}_{w_1}) log(P(\beta_{w_0}|\mathbf{C}_{w_1})). \quad (2)$$

We then have the following proposition.

Proposition 1. *Let f be a functional mapping $f : \mathbb{R}^w \to \mathbb{R}^d$ and $h_X = f(X_w)$. If f is continuous and bijective, then*

$$h_\alpha \perp\!\!\!\perp h_\beta \mid h_{\mathbf{C}} \quad \Rightarrow \quad CTE_{\alpha\to\beta|\mathbf{C}} = 0. \quad (3)$$

Please refer to Appendix C for the proof. By treating CTE as a measure of conditional independence between time series, Proposition 1 implies that

$$h_\alpha \perp\!\!\!\perp h_\beta \mid h_{\mathbf{C}} \quad \Rightarrow \quad \alpha \not\to \beta \mid \mathbf{C}. \quad (4)$$

Thus, if we can carefully design an encoder network such that all the information in the domain space is preserved in the codomain space. Next, we introduce our attention-based encoder-decoder architecture.

Attention-based Encoder-Decoder Architecture

To facilitate conditional independence test $CI(\alpha, \beta|\mathbf{C})$ where $\mathbf{C} = \{C^{(1)}, \ldots, C^{(k)}, \ldots\}$ is the conditioning set of time series, we propose an attention-based encoder-decoder framework to encode the hidden features of the time series variables into vector representations. The architectural overview of our proposed model is presented in the right half of Fig. 2. This model is structured as a sequence-to-sequence network, which takes a batch of input time series variables

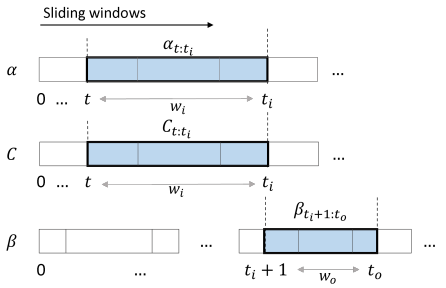


Figure 3: Data preprocessing before feeding into the encoder-decoder model, assuming conditional set C contains a single time series.

and processes them through encoding-decoding and attention layers to forecast future time steps of a target series. The input series are sequentially passed, one time window at a time. The encoder and decoder compute hidden embeddings from the inputs and the target series respectively. These are then fed into the attention layer to calculate attention scores. Subsequently, the weighted embeddings are used to compute the context vector, facilitating the prediction of the target variable by the predictor. Through time series prediction, our proposed architecture derives vector representations for all variables in the temporal data.

To simplify the representation, we first assume the causal graph has no self-loops, and then introduce a straightforward extension to accommodate scenarios where self-loops might be present. Below we elaborate on the design of the encoder-decoder model tailored for datasets devoid of self-loops.

Encoding-Decoding. We employ multi-layer long-short-term memory (LSTM) networks featuring an attention mechanism as both encoder and decoder. The encoding LSTM network f_θ^{in} takes the input time series α and C sequentially, one after another, where θ indicates the parameter of the model, as shown in Eq. (5). Similarly, the decoding network f_ψ^{out} takes the output time series β. For each triple (α, β, C) in the data $\mathcal{X}$, this network learns the nonlinear hidden representations for each time series, with the attention scores aiding in determining the significance of sequences α and C in predicting β.

We use a sliding window technique to preprocess the data (Fig. 3). The encoder processes a batch of overlapping windows, namely $\alpha_{t:(t+w_i-1)}$ and $C^{(k)}_{t:(t+w_i-1)}$, to predict the next time-step window of $\beta_{(t+w_i):(t+w_i+w_o)}$, where w_i and w_o denotes the input and output window sizes. For simplicity, we denote $(t + w_i - 1)$ as t_i and $(t + w_i + w_o)$ as t_o.

Attention. In this architecture, we have adopted the conventional scaled-dot product attention mechanism outlined in (Vaswani et al. 2017). We employed different LSTM models for the input and output sequences. For instance, considering the triple α, β, C, we compute the hidden embeddings h_α and $h_C^{(k)}$ for the input sequences α and $c^{(k)} \in C$ using the encoding LSTM f_θ^{in}. On the other hand, the hidden embedding h_β for the output sequence β is calculated using the decoding LSTM model f_ψ^{out}. Here, h_α, h_β, and $h_C^{(k)}$ belong to the space $\mathbb{R}^{h_1}$.

$$h_\alpha = f_\theta^{in}(\alpha_{t:t_i}), \quad h_C^{(k)} = f_\theta^{in}(C^{(k)}_{t:t_i}), \quad h_\beta = f_\psi^{out}(\beta_{t_i+1:t_o}) \tag{5}$$

Next, we concatenate the input hidden embeddings to obtain a matrix $h_{\alpha|C} = \{h_\alpha, h_C^{(1)}, \ldots, h_C^{(k)}, \ldots\}$, such that we have $h_{\alpha|C} \in \mathbf{R}^{j \times h_1}$ with j representing the number of input sequences. This matrix ensures that the attention scores of the inputs are collectively computed relative to the output sequence. Both $h_{\alpha|C}$ and h_β are then passed through a cross-attention module to determine the key, query, and value, as shown in Eq. (6). The weight matrices W^Q, W^K, and W^V are trainable parameters of the model, each belonging to the space $\mathbb{R}^{h_1 \times h_2}$. Consequently, the weight matrices of the cross-attention module are trained considering both inputs α and C, and updated accordingly. In this setup, h_β acts as the query input, while $h_{\alpha|C}$ serves as both key and value inputs. Thus we have $(K, V) \in \mathbf{R}^{j \times h_2}$ and $Q \in \mathbf{R}^{h_2}$.

$$Q = W^Q(h_\beta), \quad K = W^K(h_{\alpha|C}), \quad V = W^V(h_{\alpha|C}) \tag{6}$$

The cross-attention module effectively merges two distinct embedding sequences, $h_{\alpha|C}$ and h_β leveraging information from the output sequence as well. It computes attention scores for α and C relative to the output sequence β through matrix multiplication. Subsequently, the attention scores corresponding to each variable are normalized by applying a softmax function to the product of Q and K as outlined in Eq. (7). The resulting normalized attention scores are then utilized to calculate the weighted sum of the value matrix V of the inputs, employing Eq. (8), where $D \in \mathbf{R}_2^h$

$$\text{scores} = \text{softmax}(QK^T) \tag{7}$$

$$D = \text{matmul}(\text{scores}, V) \tag{8}$$

Prediction. The above equation produces a context, D, which is then fed to the prediction module to forecast the time series β of window length w_o. To generate a normalized prediction, the predictor f_ϕ incorporates a multilayer perceptron (MLP) network, featuring one fully connected layer followed by a sigmoid layer. Leveraging the context from the attention module, the predictor forecasts the output time series for the corresponding window $\beta_{t_i+1:t_o}$ as:

$$\hat{\beta}_{t_i+1:t_o} = f_\phi(D) \tag{9}$$

Training. The entire model, as described above, is trained end-to-end, in a batch-wise manner using the mean square error loss function:

$$Loss = MSE(\hat{\beta}_{t_i+1:t_o}, \beta_{t_i+1:t_o}) \tag{10}$$

The trained hidden embedding matrices h_α, h_β, and $h_C^{(k)}$ are then used for the conditional independence test as described in the following section.

Assuming self-loops in the data

We extend our encoder-decoder model to accommodate scenarios where each variable in a multivariate time series $\mathcal{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ may be dependent on itself, leading to a summary causal graph with self-loops. This adjustment

is crucial because, in such graphs, a variable is causally dependent on itself, necessitating the inclusion of its historical information in the input sequence. To achieve this, we augment our simplified encoder-decoder architecture by adding an additional input, denoted as β', which represents the past time steps of the target variable. Specifically, we increase the input dimension and conditioning set size to incorporate an extra input variable. This adjustment involves expanding the conditional set such that $C = \{C', \beta'\}$. Consequently, we include β' for predicting β using our encoder-decoder architecture. This adaptation is applied to every variable in the dataset.

In the attention module, the additional input β' is fed to the decoder LSTM f_ψ^{out}. This enables the decoder to extract the attributes of the target variable from its past and transmit them to the attention mechanism. The cross-attention module extracts the features of α, C', β' to forecast the future time points of β. As previously mentioned, we concatenate the separate hidden embeddings into a matrix as $h_{\alpha|C'\beta'} = (h_\alpha, h_{C'}, h_{\beta'})$, which are then projected to key, query, and value matrices Q, K, and V using matrix multiplication.

Hybrid Approach for Time Series Causal Discovery

Algorithm 1: Neural-HATS Algorithm

Input: $\mathcal{X}$
$M = zeros(n, n)$
foreach $\beta \in V$ **do**
  foreach $\alpha \in V \backslash \{\beta\}$ **do**
    foreach $C \in V \backslash \{\alpha, \beta\}$ **do**
      $h_\alpha, h_\beta, h_C = f_{enc-dec}(\alpha, \beta, C)$
      if $KCI(h_\alpha \perp\!\!\!\perp h_\beta \mid h_{\mathbf{C}})$ **then**
        $\mathbf{M}[\alpha, \beta] = 1$;
        break;
return: M

Most of the existing constraint-based algorithms face a challenge with exponential time complexity when discovering a causal graph. Despite efforts to enhance efficiency through various algorithms (e.g., (Absar and Zhang 2021; Assaad, Devijver, and Gaussier 2022a; Entner and Hoyer 2010; Runge et al. 2019)), many of them still suffer from high complexity. To address this, we propose a hybrid approach named Neural Hybrid Approach for Time Series causal discovery (Neural-HATS). By combining conditional independence testing with continuous optimization methods, our approach capitalizes on the accuracy of score-based techniques even with limited data while maintaining manageable time complexity. The encoder-decoder architecture, discussed above, effectively captures the hidden embeddings of each time series in the data, which are then employed to execute kernel-based conditional independence tests on every triple within the dataset. Based on this, we identify all conditional independence relations for low degrees (1 or 0) and construct a CI matrix to summarize the results, as defined below:

Definition 3 (CI Matrix)**.** *The CI matrix M is defined as:*

$$M_{\alpha,\beta} = \begin{cases} 1 & \exists C \in V \setminus \{\alpha, \beta\}, \quad \alpha \not\to \beta | C, \\ 0 & \textit{otherwise} \end{cases} \tag{11}$$

In the context of a general continuous optimization-based causal discovery method, where $\mathcal{L}$ represents the loss function, we can incorporate our CI matrix as a constraint in the following manner:

$$\begin{aligned} \min_{\mathcal{A}} \mathcal{L} &= \mathcal{L}(X, \hat{X}; \theta) + R(\theta) \\ &\text{s.t. } M_{\alpha,\beta} * \mathcal{A}_{\alpha,\beta} = 0 \quad \forall \alpha, \beta \in \mathcal{X} \end{aligned} \tag{12}$$

where $\hat{X}$ represents the reconstructed or predicted time series, θ is the set of model parameters, $\mathcal{A}$ denotes the adjacency matrix, and $R(\theta)$ is any regularization term. The CI-constraint $M_{\alpha,\beta} * \mathcal{A}_{\alpha,\beta} = 0$ requires $A_{\alpha,\beta} = 0$ when $M_{\alpha,\beta} = 1$, denoting a conditional independence relationship between α and β and the absence of a causal link between them. Inspired by the work of (Ng et al. 2022), we transform the problem into a continuous optimization framework using the augmented Lagrangian multiplier. Consequently, the CI constraint is converted into another regularization term and added to the loss function outlined in Eq. (13).

$$\min_{\mathcal{A}} \mathcal{L} = \mathcal{L}(X, \hat{X}; \theta) + R(\theta) + \lambda_{CI} ||M \circ \mathcal{A}||_F^2 \tag{13}$$

In this expression, $\circ$ denotes the Hadamard product, and λ_{CI} represents the regularization parameter, serving to balance the impact of the CI matrix. The term $||M \circ \mathcal{A}||_F^2$ plays an important role in diminishing the influence of the penalized elements in $\mathcal{A}$ by M. Consequently, solving Eq. (13) using any continuous optimization method enables the inference of a more precise summary causal graph $\mathcal{A}$. The incorporation of this information subsequently diminishes the required sample size for continuous optimization. Additionally, in scenarios involving large and sparse graphs, the use of CI tests facilitates the removal of more edges, thereby elevating the overall efficacy of causal discovery.

The proposed algorithm for generating the CI matrix M is outlined in Algorithm 1. Initially, M is initialized with zeros, assuming dependencies between all variable pairs. Then, a conditional independence test is conducted between every β and α in V for each $C \in \mathbf{V} \backslash \{\alpha, \beta\}$, with the condition that $\alpha \neq \beta$. The KCI test is executed on the hidden embeddings obtained from our proposed attention-based encoder-decoder model. We consider the following hypotheses:

$$H_0 : h_\alpha \perp\!\!\!\perp h_\beta \mid h_C \tag{14}$$

$$H_1 : h_\alpha \not\perp\!\!\!\perp h_\beta \mid h_C \tag{15}$$

Here, h_α, h_β, and h_C denote the hidden embeddings of time series α, β, and C respectively, and $f_{enc-dec}$ represents the computations of f_θ^{in} and f_ψ^{out}. The KCI test uses the nonlinear representation of the time series extracted from the trained encoder as inputs, ensuring the incorporation of both historical information and hidden features in the conditional

independence assessment. If H_0 is not rejected for any C, indicating independence between α and β given C, the corresponding element in the CI-matrix $M_{\alpha,\beta}$ is set to 1, and the test is repeated for the next triple.

Experiments

In this section, we conduct extensive experiments to evaluate Neural-HATS using different data. We will first evaluate its performance of conducting conditional independence tests alone. Then, we will evaluate its performance of causal discovery by incorporating it with various base models.

Dataset

Synthetic Data. For the generation of multivariate time series data, we employ the vector autoregressive method, given by the equation:

$$X_t = A^T \sum_{j=1}^{5} \beta_j \cos(X_{t-j} + 1) + \epsilon \tag{16}$$

Here, X_t signifies a vector of d variables at time step t, β is the regression coefficient, and ϵ represents standard Gaussian noise. The noise scale is kept below 1 and is proportional to the value of n. The non-linear relationship between time series is introduced through the $\cos$ function. The adjacency matrix A of the underlying causal graph is generated using the Erdős–Rényi model (Newman 2018). We employ a maximum lag of 5 to generate the causal relationships and initialize the values $\mathbf{X}_{0:4}$ randomly. We have generated multiple datasets in this way, each possessing a distinct underlying causal graph A.

Real Data. We also evaluate the performance of our Neural-HATS and other baselines using real-world Netsim dataset (Smith et al. 2011). This dataset comprises realistic simulated functional magnetic resonance imaging (fMRI) time series data representing the blood-oxygen-level-dependent (BOLD) signals across various regions of the human brain. The underlying connectivity in this dataset reflects the causal relationships between different brain regions, with the adjacency matrix depicting this relationship and the nodes representing the various brain regions. The Netsim dataset contains simulations from numerous areas of the brain. From that, we select the fifth simulation, Sim-5.mat, for our analysis. This particular simulation consists of data with 5 nodes and 1200 time steps for 50 participants. We opted for this simulation due to its ample data length, which is conducive to training our deep-learning models.

Experimental Setting

We randomly sample 500 consecutive data points from each dataset to construct the CI matrix using our Neural-HATS algorithm. Following this, the entire dataset is used to train the baseline methods based on continuous optimization. For the CI matrix construction, we perform order-1 CI tests (i.e., $|C| = 1$) on the synthetic data and order-0 (marginal) CI tests (i.e., $C = \emptyset$) on the Netsim dataset. In order to perform marginal tests $CI(\alpha, \beta|\emptyset)$, we propose to first manually construct a time series C with random Gaussian noise and then perform the CI test for $CI(\alpha, \beta|C)$ as usual.

The encoder-decoder architecture is designed with three LSTM layers for synthetic data, with $h_1 = 32$ and a dropout of 0.2. For real data, a different encoder-decoder setup with two LSTM layers and dimension $h_1 = 50$ is employed. Input and output window sizes are set to $w_i = 10$ and $w_o = 5$ for synthetic data, while for real data, $w_i = 5$ and $w_o = 1$ are used for preprocessing before feeding into the LSTM. The model is trained with a learning rate of 1×10^{-3} and a batch size of 128 to update h_α, h_β, and h_C for all data instances. For the attention layers, we use $h_2 = 2 * h_1$. Following this, we perform the KCI test with a significance level of 0.1. Our model is evaluated across various synthetic datasets and a real-world dataset. All experiments are conducted in PyTorch and run on a computer with Ubuntu 20.04.4 LTS, featuring an Intel(R) Core(TM) i9-10900X CPU and NVIDIA GeForce RTX 3080 10GB GPU. The codes for replicating all our experiments are available at `https://github.com/SaimaAbsar/Neural-HATS`.

Base Models

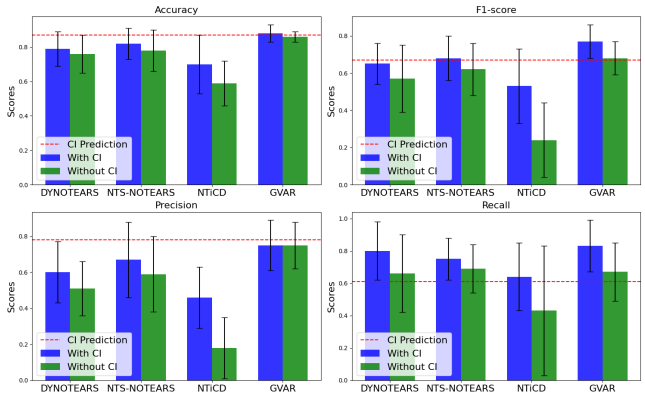
Neural-HATS stands out as a hybrid method, providing a generated CI matrix that can be seamlessly incorporated as a regularization term into any base continuous optimization model for temporal causal discovery. In this study, we employ four algorithms as the base models, namely DYNOTEARS (Pamfil et al. 2020), NTS-NOTEARS (Sun et al. 2021), NTiCD (Absar, Wu, and Zhang 2023), and GVAR (Marcinkevičs and Vogt 2021), chosen for their continuous score-based nature. We also consider these methods as baselines against which we can assess the improvements offered by Neural-HATS.

Note that as DYNOTEARS and NTS-NOTEARS generate window causal graphs, we transform them into a summary graph by considering an edge from one variable to another if there exists an edge with any lag between the corresponding variables in the window causal graph.

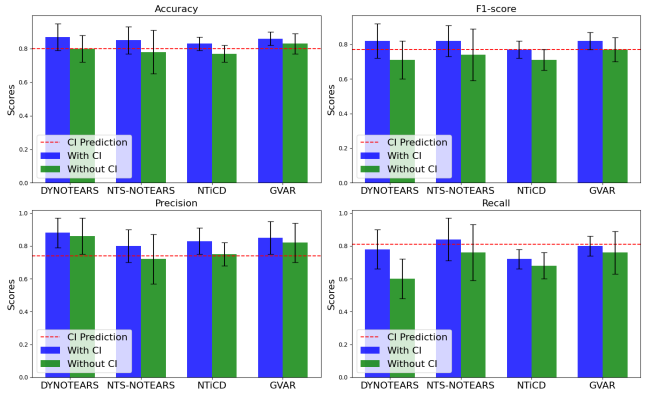
Results

Conditional Independence Testing We first present the results of conditional independence tests conducted on both synthetic ER data and real-world Netsim data in Table 1. The table shows the mean and standard deviation across 10 simulations involving different causal graphs. Here, precision denotes the ratio of the pairs that are genuinely independent among all that are detected independent. On the other hand, recall indicates the proportion of correctly identified independent pairs among all independent pairs in the ground truth. Accuracy reflects the percentage of correctly detected independent and also dependent pairs. Notably, Neural-HATS demonstrates high accuracy in detecting independent pairs within the dataset, as evident from the table.

Hybrid Causal Discovery The performance of the hybrid method offered by Neural-HATS using all four base models is illustrated in the plot in Fig. 4. Here, the results for the four base models before and after incorporating Neural-HATS are presented. Fig. 4a displays the performance results for synthetic data, while Fig. 4b shows the results for the Netsim dataset. The metric values are averaged across 10



(a) Synthetic data.



(b) Real data (Netsim).

Figure 4: The performance of different baselines w/ and w/o the incorporation of Neural-HATS.

experiment runs and shown along with their variances. For each experiment, we tuned the regularization parameter λ_{CI} in the logarithmic scale from 0.05 to 500 for all methods and recorded the highest performance obtained. The plots in the figure compare the results obtained from the base models with the hybrid methods, which show that Neural-HATS generally outperforms all base models in terms of accuracy, precision, recall, and f1-score.

Additionally, we present the performance obtained by converting the CI matrix M produced by Neural-HATS directly to the adjacency matrix A with red dashed lines in each plot. This conversion simply involves changing the 1's in M to 0's in A as 1 in M represents independence. Here, we only present the mean metric values for visualization efficiency, as the standard deviation typically ranges from 0.04 to 0.18. The size of the conditioning set is 1 for the synthetic data and 0 for the real data. The results from both real and synthetic data show that the CI prediction falls short of the best version of Neural-HATS in all settings. This demonstrates the advantage of Neural-HATS as a hybrid method that leverages the strengths of both conditional independence testing and continuous optimization. This result also implies that low-order CI tests are sufficient to achieve high causal discovery accuracy.

Table 1: Independent test results for Synthetic ER data and Real data, where the metrics show the mean values of 10 experiment runs.

	Synthetic data	Real data
Accuracy	0.81 ±0.015	0.76 ±0.075
F1-score	0.81 ±0.025	0.83 ±0.066
Precision	0.75 ±0.079	0.87 ±0.040
Recall	0.88 ±0.057	0.80 ±0.103

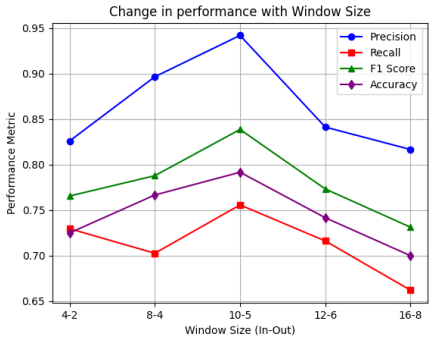


Figure 5: Performance variation of proposed conditional independence test with respect to different input-output window sizes on synthetic data.

Influence of Window Size As implied by Proposition 1, the conditional independence relations in the embedding space are maintained in the time series space if the encoder network can preserve the information in the domain space in the codomain space. In this experiment, we fix the output dimension of the encoding LSTM to 32 and vary the window sizes of the input and output sequences in order to study the impact of the varying windows on the performance of CI testing. The results are shown in Fig. 5 for synthetic data where the X-axis represents the input-output window sizes. As can be seen, utilizing an input window length of 10 to predict the output with a window of 5 in the encoder-decoder model yields the highest performance metrics in terms of F1-score, accuracy, precision, and recall. This result shows the importance of carefully designing the encoder network as indicated by our theoretical results.

Conclusion

In this paper, we introduced Neural-HATS, a novel method that merges conditional independence (CI) tests with continuous optimization for temporal causal discovery. We proposed an attention-based encoder-decoder framework that extracts vector embeddings from time series data, facilitating CI tests using kernel-based conditional independence (KCI) testing. By conducting lower-order CI tests on these vectors, we constructed a CI matrix, which is then integrated into a continuous optimization-based causal discovery method. We demonstrated performance improvement by incorporating Neural-HATS into four state-of-the-art score-based methods using synthetic and real-world data.

Acknowledgments

This work was supported in part by NSF 1910284 and 2142725.

References

Absar, S.; Wu, Y.; and Zhang, L. 2023. Neural Time-Invariant Causal Discovery from Time Series Data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Absar, S.; and Zhang, L. 2021. Discovering Time-invariant Causal Structure from Temporal Data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2807–2811.

Assaad, C. K.; Devijver, E.; and Gaussier, E. 2022a. Entropy-Based Discovery of Summary Causal Graphs in Time Series. *Entropy*, 24(8): 1156.

Assaad, C. K.; Devijver, E.; and Gaussier, E. 2022b. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research*, 73: 767–819.

Assaad, K.; Devijver, E.; Gaussier, E.; and Ait-Bachir, A. 2021. A mixed noise and constraint-based approach to causal inference in time series. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, 453–468. Springer.

Cheng, Y.; Yang, R.; Xiao, T.; Li, Z.; Suo, J.; He, K.; and Dai, Q. 2023. CUTS: Neural Causal Discovery from Irregular Time-Series Data. *arXiv preprint arXiv:2302.07458*.

Chu, Y.; Wang, X.; Ma, J.; Jia, K.; Zhou, J.; and Yang, H. 2020. Inductive granger causal modeling for multivariate time series. In *2020 IEEE International Conference on Data Mining (ICDM)*, 972–977. IEEE.

Entner, D.; and Hoyer, P. O. 2010. On causal discovery from time series data using FCI. *Probabilistic graphical models*, 121–128.

Gerhardus, A.; and Runge, J. 2020. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33: 12615–12625.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Granger, C. W. 1980. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2: 329–352.

Granger, S. 2001. Social engineering fundamentals, part I: hacker tactics. *Security Focus, December*, 18.

Guo, R.; Cheng, L.; Li, J.; Hahn, P. R.; and Liu, H. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4): 1–37.

Hasan, U.; Hossain, E.; and Gani, M. O. 2023. A Survey on Causal Discovery Methods for Temporal and Non-Temporal Data. *arXiv preprint arXiv:2303.15027*.

Hiemstra, C.; and Jones, J. D. 1994. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5): 1639–1664.

Hyvärinen, A.; Zhang, K.; Shimizu, S.; and Hoyer, P. O. 2010. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5).

Krich, C.; Runge, J.; Miralles, D. G.; Migliavacca, M.; Perez-Priego, O.; El-Madany, T.; Carrara, A.; and Mahecha, M. D. 2020. Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach. *Biogeosciences*, 17(4): 1033–1061.

Li, Y.; Xia, R.; Liu, C.; and Sun, L. 2022. A hybrid causal structure learning algorithm for mixed-type data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7435–7443.

Löwe, S.; Madras, D.; Zemel, R.; and Welling, M. 2022. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, 509–525. PMLR.

Luo, L.; Liu, W.; Koprinska, I.; and Chen, F. 2015. Discovering causal structures from time series data via enhanced granger causality. In *AI 2015: Advances in Artificial Intelligence: 28th Australasian Joint Conference, Canberra, ACT, Australia, November 30–December 4, 2015, Proceedings 28*, 365–378. Springer.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Malinsky, D.; and Spirtes, P. 2018. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, 23–47. PMLR.

Marcinkevičs, R.; and Vogt, J. E. 2021. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*.

Mogensen, S. W.; Hansen, N. R.; et al. 2020. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1): 539–559.

Moraffah, R.; Sheth, P.; Karami, M.; Bhattacharya, A.; Wang, Q.; Tahir, A.; Raglin, A.; and Liu, H. 2021. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63: 3041–3085.

Nauta, M.; Bucur, D.; and Seifert, C. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1): 19.

Newman, M. E. 2018. Estimating network structure from unreliable measurements. *Physical Review E*, 98(6): 062321.

Ng, I.; Lachapelle, S.; Ke, N. R.; Lacoste-Julien, S.; and Zhang, K. 2022. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, 8176–8198. PMLR.

Pamfil, R.; Sriwattanaworachai, N.; Desai, S.; Pilgerstorfer, P.; Georgatzis, K.; Beaumont, P.; and Aragam, B. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, 1595–1605. PMLR.

Peters, J.; Janzing, D.; and Schölkopf, B. 2013. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26.

Runge, J. 2018. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7): 075310.

Runge, J. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, 1388–1397. Pmlr.

Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; and Sejdinovic, D. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11): eaau4996.

Shojaie, A.; and Fox, E. B. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9: 289–319.

Smith, S. M.; Miller, K. L.; Salimi-Khorshidi, G.; Webster, M.; Beckmann, C. F.; Nichols, T. E.; Ramsey, J. D.; and Woolrich, M. W. 2011. Network modelling methods for FMRI. *Neuroimage*, 54(2): 875–891.

Sugihara, G.; May, R.; Ye, H.; Hsieh, C.-h.; Deyle, E.; Fogarty, M.; and Munch, S. 2012. Detecting causality in complex ecosystems. *science*, 338(6106): 496–500.

Sun, X. 2008. Assessing nonlinear Granger causality from multivariate time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 440–455. Springer.

Sun, X.; Schulte, O.; Liu, G.; and Poupart, P. 2021. NTS-NOTEARS: Learning Nonparametric DBNs With Prior Knowledge. *arXiv preprint arXiv:2109.04286*.

Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; and Fox, E. B. 2021. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4267–4279.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2022. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4): 1–36.

Zanga, A.; Ozkirimli, E.; and Stella, F. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151: 101–129.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Appendix A: Related Work

Granger causality (Granger 1969; Shojaie and Fox 2022), a key approach for time series causal discovery, has evolved from its origins in linear processes to encompass non-linear relationships through various extensions (Löwe et al. 2022; Lütkepohl 2005; Sugihara et al. 2012; Sun 2008; Tank et al. 2021). For example, the Temporal Causal Discovery Framework (TCDF) (Nauta, Bucur, and Seifert 2019) utilizes CNNs and attention mechanisms, while InGRA (Chu et al. 2020) applies LSTM-based attention for causal inference. Other methods, such as Tank et al.'s (Tank et al. 2021) sparse neural network approach and Cheng et al.'s (Cheng et al. 2023) CUTS algorithm, push the boundaries of Granger causality into non-linear domains. Constraint-based algorithms infer causal relationships by testing conditional independence (CI), as seen in PCMCI (Runge et al. 2019), which adapts to both linear and nonlinear data. Extensions such as PCMCI+, LPCMCI, tsFCI, and FCITMI have been developed to enhance scalability (Assaad, Devijver, and Gaussier 2022a; Entner and Hoyer 2010; Gerhardus and Runge 2020; Runge 2020), but often assume stationarity and face dimensionality challenges. Recent developments like μ-PC (Absar and Zhang 2021) use neural networks for CI measures, extending causal discovery to time-dependent domains, despite limitations like a focus on Recurrent Marked Temporal Point Processes (RMTPP). Score-based approaches, which optimize predefined score functions to learn Bayesian networks, have also advanced significantly with deep learning integration. Methods such as GVAR (Marcinkevičs and Vogt 2021), DYNOTEARS (Pamfil et al. 2020), NTS-NOTEARES (Sun et al. 2021), and NTiCD (Absar, Wu, and Zhang 2023) have extended these approaches to time series data to detect Granger causality. Despite their computational robustness, these methods often require large datasets to accurately determine causal graphs.

To address existing limitations, hybrid causal discovery algorithms like NBCB (Assaad et al. 2021), HCM (Li et al. 2022), and SVAR-GFCI (Malinsky and Spirtes 2018) combine different frameworks, enhancing causal discovery in multivariate time series. Despite their innovations, these methods still face challenges such as high time complexity and restrictions to DAGs, which may not always be suitable for real-world temporal data. Structural equation models (SEM), used by approaches like TiMINo (Peters, Janzing, and Schölkopf 2013) and VarLinGAM (Hyvärinen et al. 2010), continue to play a significant role in causal discovery, although they are often limited to linear data and can struggle with larger datasets.

Appendix B: Kernel-based Conditional Independence Test

The centralized kernal matrices $\tilde{\mathbf{K}}_X$ of the sample $\mathbf{x}$ could be constructed by $\tilde{\mathbf{K}}_X = \mathbf{H}\mathbf{K}_X\mathbf{H}$, where the (i,j)th entry of $\mathbf{K}_X$ is $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_X^2})$, and σ_X denotes the kernel width. Similar notations are used for Y and Z. The authors in (Zhang et al. 2012) further construct the centralized kernel matrices $\tilde{\mathbf{K}}_{\ddot{X}|Z}$ and $\tilde{\mathbf{K}}_{Y|Z}$ that corresponds

to the functions $\tilde{f}(\ddot{X})$ and $\tilde{g}'$. The authors then show that conditional independence holds if and only if the functions in certain kernel spaces are uncorrelated, formally expressed as:

$$X \perp\!\!\!\perp Y|Z \iff \mathbf{E}[\tilde{f}g = 0],\ \forall \tilde{f} \in \mathcal{E}_{XZ}\ and\ \tilde{g}' \in \mathcal{E}'_{YZ} \quad (17)$$

Based on kernel ridge regression and EVD decomposition they derived an equivalent condition for conditional independence based on the kernel matrix, which could be expressed as:

$$X \perp\!\!\!\perp Y|Z \iff T_{CI} \overset{d}{=} \hat{T}_{CI} \quad (18)$$

T_{CI} and $\hat{T}_{CI}$ denote the test statistic and its asymptotic distribution:

$$T_{CI} \triangleq \frac{1}{n} Tr(\tilde{\mathbf{K}}_{\ddot{X}|Z} \tilde{\mathbf{K}}_{Y|Z}) \quad (19)$$

$$\hat{T}_{CI} \triangleq \frac{1}{n} \sum_{k=1}^{n^2} \dot{\lambda}_k \cdot z_k^2 \quad (20)$$

where $\dot{\lambda}_k$ are eigenvalues of $\hat{\mathbf{w}}\hat{\mathbf{w}}^\intercal$. The mean and variance of $\hat{T}_{CI}$ under null hypothesis $X \perp\!\!\!\perp Y|Z$, on the given sample $\mathcal{D}$, is:

$$\mathbb{E}[\hat{T}_{CI}|\mathcal{D}] = \frac{1}{n} Tr(\hat{\mathbf{w}}\hat{\mathbf{w}}^\intercal) \quad (21)$$

$$\mathbb{V}ar[\hat{T}_{CI}|\mathcal{D}] = \frac{2}{n^2} Tr[(\hat{\mathbf{w}}\hat{\mathbf{w}}^\intercal)^2] \quad (22)$$

Subsequently, an empirical null distribution under the null hypothesis is simulated by drawing random samples from the χ^2 distribution. Finally, the p-value is calculated as the probability of the simulated distribution exceeding the test statistic.

Appendix C: Proof of Proposition 3

Proof. By definition, we have

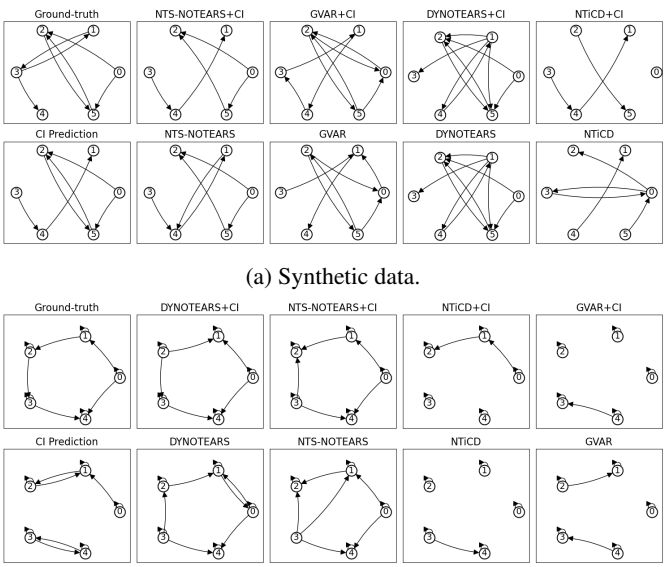
$$\begin{aligned} CTE_{\alpha\to\beta|\mathbf{C}} = &\sum_{\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}} P(\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}) \log\left(P(\beta_{w_0} \mid \alpha_{w_1},\mathbf{C}_{w_1})\right) \\ &- \sum_{\beta_{w_0},\mathbf{C}_{w_1}} P(\beta_{w_0},\mathbf{C}_{w_1}) \log\left(P(\beta_{w_0} \mid \mathbf{C}_{w_1})\right) \end{aligned} \quad (23)$$

Since f is continuous and bijective, the conditional independence relationship in the codomain space is maintained in the domain space. Thus, we have:

$$h_\alpha \perp\!\!\!\perp h_\beta \mid h_{\mathbf{C}} \quad \Rightarrow \quad \alpha_{w_1} \perp\!\!\!\perp \beta_{w_0} \mid C_{w_1} \quad (24)$$

It follows that $P_{\beta_{w_0}|\alpha_{w_1},C_{w_1}} = P_{\beta_{w_0}|C_{w_1}}$ and $P_{\alpha_{w_1}\beta_{w_0}|C_{w_1}} = P_{\alpha_{w_1}|C_{w_1}} P_{\beta_{w_0}|C_{w_1}}$. The first term of Eq. (23) can be rewritten as:

$$\begin{aligned} &\sum_{\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}} P(\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}) \log(P(\beta_{w_0} \mid \alpha_{w_1},\mathbf{C}_{w_1})) \\ =&\sum_{\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}} P(\beta_{w_0},\alpha_{w_1} \mid \mathbf{C}_{w_1}) P(\mathbf{C}_{w_1}) \log(P(\beta_{w_0} \mid \mathbf{C}_{w_1})) \\ =&\sum_{\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}} P(\alpha_{w_1} \mid \mathbf{C}_{w_1}) P(\beta_{w_0} \mid \mathbf{C}_{w_1}) P(\mathbf{C}_{w_1}) \log(P(\beta_{w_0} \mid \mathbf{C}_{w_1})) \\ =&\sum_{\beta_{w_0},\alpha_{w_1},\mathbf{C}_{w_1}} P(\alpha_{w_1} \mid \mathbf{C}_{w_1}) P(\beta_{w_0},\mathbf{C}_{w_1}) \log(P(\beta_{w_0} \mid \mathbf{C}_{w_1})) \\ =&\sum_{\beta_{w_0},\mathbf{C}_{w_1}} P(\beta_{w_0},\mathbf{C}_{w_1}) \log(P(\beta_{w_0} \mid \mathbf{C}_{w_1})) \end{aligned} \quad (25)$$



(a) Synthetic data.

(b) Real data (Netsim).

Figure 6: Comparing the causal graphs obtained directly from the CI matrix and both hybrid and raw baselines.

By plugging the results to Eq. (23) we have $CTE_{\alpha\to\beta|\mathbf{C}} = 0$. □

Appendix D: Examples of Hybrid Causal Discovery

Causal graphs predicted by different baseline methods are depicted in Fig. 6, arranged from the highest to lowest F1-scores. This figure compares the causal graphs predicted by the raw baselines with those generated by the hybrid methods for both synthetic and real datasets. Additionally, we present the causal graph obtained directly from the CI matrix (bottom-left). Compared with the ground truth graph (top-left), we can see that the hybrid methods yield the most accurate summary causal graphs for these datasets. For instance, in the synthetic data, the graph predicted by raw GVAR includes a redundant relationship from node 0 to 1, which is removed by the CI constraint in the hybrid GVAR+CI method. We observe a similar redundant edge removal in NTS-NOTEARS from nodes 1 to 4. We also observe the discovery of true connections in the hybrid method that were not detected by the raw baseline, for example, the edge from node 3 to 4 in NTiCD+CI and the self-loop of node 3 in the real data by NTS-NOTEARS+CI. For the real data, the false causal connection from node 1 to 0 by DYNOTEARS is removed by the hybrid DYNOTEARS+CI method, resulting in a more precise graph.