Learning Flexible Time-windowed Granger Causality Integrating Heterogeneous Interventional Time Series Data

Ziyi Zhang
zyzhang@tamu.edu
Texas A&M University
College Station, Texas, USA

Shaogang Ren
shaogang@tamu.edu
Texas A&M University
College Station, Texas, USA

Xiaoning Qian
xqian@tamu.edu
Texas A&M University, College Station, Texas
Brookhaven National Laboratory, Upton, New York
USA

Nick Duffield
duffieldng@tamu.edu
Texas A&M University
College Station, Texas, USA

ABSTRACT

Granger causality, commonly used for inferring causal structures from time series data, has been adopted in widespread applications across various fields due to its intuitive explainability and high compatibility with emerging deep neural network prediction models. To alleviate challenges in better deciphering causal structures unambiguously from time series, the use of interventional data has become a practical approach. However, existing methods have yet to be explored in the context of imperfect interventions with unknown targets, which are more common and often more beneficial in a wide range of real-world applications. Additionally, the identifiability issues of Granger causality with unknown interventional targets in complex network models remain unsolved. Our work presents a theoretically-grounded method that infers Granger causal structure and identifies unknown targets by leveraging heterogeneous interventional time series data. We further illustrate that learning Granger causal structure and recovering interventional targets can mutually promote each other. Comparative experiments demonstrate that our method outperforms several robust baseline methods in learning Granger causal structure from interventional time series data.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Mathematics of computing** → **Causal networks**.

KEYWORDS

Granger causality; Causal structure learning; Interventional time series data

ACM Reference Format:
Ziyi Zhang, Shaogang Ren, Xiaoning Qian, and Nick Duffield. 2024. Learning Flexible Time-windowed Granger Causality Integrating Heterogeneous Interventional Time Series Data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3672023

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3672023

1 INTRODUCTION

Time series data, capturing complex systems dynamic behaviors, are widely collected in many research areas, such as economics, bioinformatics, and geo-informatics. Due to the rapid advancements in sensor and computing technologies, there has been a significant increase in research modeling time series data in recent years. Researchers have developed various methods leveraging time series data to perform related analysis such as optimization [29, 30], classification [26, 44, 46, 47], clustering [19, 27, 48], forecasting [41, 49], and causal structure learning [12, 21, 24, 28, 31, 39]. Among these tasks, causal structure learning is particularly challenging but important. Multivariate time series data, which capture the evolving states of multiple variables over time, facilitate deriving better systems understanding across various domains. Causal structure learning in multivariate time series data focuses on understanding how different variables influence each other. This knowledge is beneficial for explaining the data generation process and guiding the design of time series analysis methods [14].

Granger causality has been widely used for analyzing time series data to discover causal relationships in numerous real-world applications, including modern healthcare systems [42], medical time series generation [23] and time series anomaly detection [35]. Many methods for learning causal structures in time series have been developed based on the principles of Granger causality [6, 21, 28, 39, 43]. However, Granger causality tests based on linear models can be ineffective when faced with even slight non-linear causal relationships in the measurements. Consequently, a significant amount of research efforts have been focused on addressing issues for Granger causality considering non-linearities [21, 28, 39].

Learning causal structures based solely on observational data is challenging [36, 37] because, under the faithfulness assumption, the true causal structure can only be identified within a Markov Equivalence Class (MEC) [40]. However, this identifiability improves when we consider interventional data. We have observed that domain experts might be able to gather interventional data in practice, where the underlying generative process varies across different conditions. This characteristic of distribution shift presents unique challenges as well as opportunities for learning causal structures in time series.

In these scenarios, the causal structure can be identified within an Interventional Markov Equivalence Class ($\mathcal{I}$-MEC), which is a more specific subset of the Markov equivalence class [2, 17, 45]. With sufficient interventional observations, the causal structure can be precisely identified [9, 10]. Numerous methods have approached causal structure learning with interventional data by framing it within a continuous optimization framework [2, 12, 31], incorporating a continuous acyclicity constraint [50]. To address identifiability challenges with time series data, the authors of [12] have extended the work of [2, 31], to effectively handle observational and interventional time series data under both perfect [11, 45] and imperfect interventions [34] with known interventional targets [2] (See Figure 1). However, in real-world applications, imperfect interventions with unknown targets are more common [51], requiring information about interventional targets limits their applications in more general cases. Therefore, learning causal structures from interventional time series data is still an open problem.

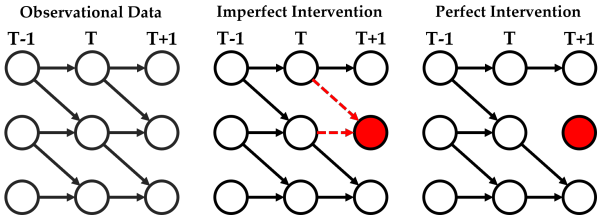

Figure 1: Intervention types on time series: With known interventional target (red nodes), altered all causal relationships from parent nodes in imperfect interventions (red dotted lines) versus disconnection from parent nodes in perfect interventions.

In this paper, we emphasize on the following parts, compared to previous work:

- **Practicality.** Most methods require knowledge regarding interventional targets. However, in practical scenarios, distinguishing which variables originate from the non-intervened domain and identifying the exact interventional targets often proves to be challenging.
- **Accuracy.** In previous research, understanding of imperfect interventions has been limited to the node level. However, as illustrated in Figure 2, edge-level imperfect interventions can clarify the specific situations leading to imperfect interventions, which has not been explicitly studied.
- **Identifiability.** Despite the development of advanced Score-based [12, 24, 31] or Granger causality-based [6, 21, 28, 39, 43] causal structure learning methods for time series, issues related to the identifiability of Granger causality with unknown interventional targets remain unresolved.

Consequently, we introduce a theoretically-guaranteed Interventional Granger Causal structure learning (*IGC*) method. This approach is designed for the simultaneous inference of Granger causal structure and the identification of unknown interventional targets at the edge level. It also leverages interventional time series data across multiple domains, efficiently differentiating among those

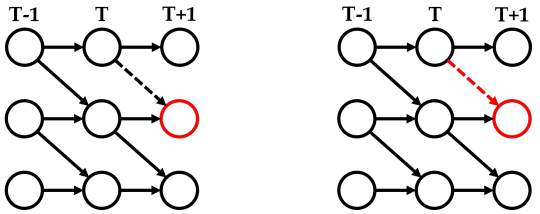

Figure 2: (Left): Existing methods (node-level imperfect intervention on an unknown target) can only identify the exact node(s); (Right): whereas our method (edge-level intervention identification) can identify both the node(s) and exact edge(s).

that have not been intervened upon and those that have, especially in scenarios where interventional targets are unknown and the distinctions are not readily apparent. In summary, the main contributions of the paper are highlighted as:

- We have formalized the task of learning Granger causal structure from heterogeneous interventional time series data. The interventional targets are unknown and samples from observational distribution may be indistinguishable from other interventional distribution.
- A theoretically-guaranteed method called Interventional Granger Causal structure learning (*IGC*) is developed to simultaneously infer Granger causal structure and identify unknown interventional targets at the edge level.
- We have shown that the exact minimization of the proposed objective will identify the $(\mathcal{I}, \mathcal{D})$-Markov equivalence class of the ground truth graph in the context of unknown target setting, then resolve the identifiability issues of Granger causality.
- Extensive experiments on both synthetic and real-world time series data have demonstrated our proposed *IGC* outperforms several robust baselines by utilizing interventional data.

2 RELATED WORK

Granger Causal Structure Learning: Much work has been conducted on inferring causal structure based on Granger causality in multivariate time series. Recent approaches for inferring Granger causal structure leverage the expressive power of neural network and are often based on regularized autoregressive models. [1] proposed the Lasso Granger method. [39] proposed the sparse-input multi-layer perceptron (MLP) and long short-term memory (LSTM) to model the nonlinear Granger causality within multivariate time series. [21] integrated an efficient economy statistical recurrent unit architecture with input layer wights regularized in a group-wise manner. [28] proposed a generalized vector autoregression model that utilizes self-explaining neural networks (SENNs) for inferring Granger causal structure, with an additional focus on detecting signs of Granger-causal effects. [4, 5] proposed a Granger causal discovery algorithm that builds a causal adjacency matrix for imputed and high-dimensional data using sparse regularization. Although these methods are powerful techniques for inferring Granger causal structure, they do not fully utilize interventional data, nor do they address the identifiability problem.

Causal Structure Learning from Interventional Data: A series of parametric studies treat data from different distributions, often referred to as domains or environments, as interventional data. [13] studied the problem of causal structure learning in linear systems from observational data given in multiple domains, across which the causal coefficients may vary. [45] studied the problem of causal structure learning in the setting where both observational and interventional data is available and extended the identifiability results from perfect intervention [17] to general interventions. [2] proposed a differentiable causal structure learning method for static data that can leverage perfect, imperfect and unknown target interventions using score function to identify the $\mathcal{I}$-MEC. [24] propose a novel latent intervened non-stationary learning method to recover the domain indexes and the causal structure. [12] extends [2, 31] to address both observational and interventional time series data, including perfect and imperfect interventions with known targets. However, effectively handling both observational and interventional time series data in an imperfect setting with unknown interventional targets remains a challenge.

3 PRELIMINARIES

Non-linear Granger Causality: Consider multivariate time series $\mathcal{T} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, where $\mathbf{x} \in \mathbb{R}^d$. Assume that causal relationships between variables are given by the following structural model:

$$x^i_{t+1} = g_i(x^1_{1:t}, \ldots, x^d_{1:t}) \quad \text{for } 1 \le i \le d, \tag{1}$$

where $g_i(\cdot)$ is a function that specifies how the past values are mapped to series i. Time series j is *Granger non-causal* for time series i if for all $x^1_{1:t}, \ldots, x^d_{1:t}$ and all $x^j_{1:t} \neq \hat{x}^j_{1:t}$ [39]:

$$g_i(x^1_{1:t}, \ldots, x^j_{1:t}, \ldots, x^d_{1:t}) = g_i(x^1_{1:t}, \ldots, \hat{x}^j_{1:t}, \ldots, x^d_{1:t}). \tag{2}$$

Interventions: In the context of causal structure learning, an intervention on a variable x_i, involves altering its conditional probability $\mathbf{P}(x_i|\mathbf{PA}(x_i))$ to a new conditional probability $\widetilde{\mathbf{P}}(x_i|\mathbf{PA}(x_i))$, where $\mathbf{PA}(x_i)$ is the set of parents of the node x_i in the causal graph. It is possible to apply interventions to several variables at once. The set of variables on which k-th interventions are made is referred to as the *interventional targets*, symbolized by $\mathbf{I}_k \in \mathbb{R}^d$. The *interventional family* is defined as $\mathcal{I} := (\mathbf{I}_1, \ldots, \mathbf{I}_n)$, where n represents the total number of interventions conducted.

Types of Interventions: The type of interventions depicted in Figure 1 are generally categorized as imperfect intervention (also known as soft or parametric intervention) [8, 34]. In contrast, a specific case within this broad category is the perfect intervention (also referred to as hard or structural intervention), where $\mathbf{P}(x_i|\mathbf{PA}(x_i)) = \mathbf{P}(x_i)$ [2, 11, 22, 45].

4 INTERVENTIONAL GRANGER CAUSAL STRUCTURE LEARNING

In this section, we first discuss the challenge of learning Granger causal structure from interventional time series data in situations where the interventional targets are unknown, and samples from the observational distribution is indistinguishable from those of other interventional distributions. Subsequently, we propose our Interventional Granger Causal structure learning (*IGC*) method to learn both the underlying Granger causal structure and unknown interventional targets across different environments.

4.1 Granger Causality with Interventions

First, we start with a linear Lasso Granger methodology capable of handling both observational and interventional data, drawing inspiration from the concepts applied to independent and identically distributed (i.i.d.) datasets [2] and structural vector autoregression model [12]. The core principle of these methods involve constructing a Directed Acyclic Graph (DAG) representing the ground truth causal graph from the interventional data. This is achieved by incorporating a distinct distribution family specifically for the intervened nodes within the log-likelihood objective. Unlike these methods that model the post-intervention distribution, our approach concentrates on comparing the distributions before and after the intervention under the framework of Granger causality to help interpret the impact of the intervention on time series data more clearly. Specifically, we employ $\mathbf{W}_{e_0} \in \mathbb{R}^{d\times d}$ to represent the parameters of the density function for observational data under the condition of no interventions. For each intervention $\mathbf{I}_k \in \mathcal{I}$, we define another corresponding set of parameters $\mathbf{W}_{e_k} \in \mathbb{R}^{d\times d}$, which captures the differences in density functions before and after the k-th intervention. In other words, the density function after the k-th intervention can be represented as $\mathbf{W}_{e_0} + \mathbf{W}_{e_k}$. The collection of these parameters is denoted by $\boldsymbol{W} := \{\mathbf{W}_{e_0}, \mathbf{W}_{e_1}, ..., \mathbf{W}_{e_n}\}$. The integrated training loss function, as described in Equation (3), takes into account both observational and interventional data:

$$\mathcal{L}(\mathbf{X}; \mathbf{W}) = \sum_{k=1}^{n} \sum_{t=1}^{T} \sum_{\tau=1}^{l} \mathcal{L}_k(\mathbf{X}_t - (\mathbf{W}_{e_0} + \mathbf{W}_{e_k})\mathbf{X}_{t-\tau}), \tag{3}$$

where n represents the total number of interventions and $\mathcal{L}_k$ signifies the training loss based on the time series data from the k-th intervention. In this model, we do not know which environment among n environments is non-intervend and we assume that $\mathbf{W}_{e_0}$ remains constant across n different environments, interventions, domains, or distributions. A time series j is *Granger non-causal* for time series i if and only if the corresponding weight $\mathbf{w}_{ij}$ in the matrix $\mathbf{W}_{e_0}$ is zero. Intuitively, if all elements in $\mathbf{W}_{e_k}$ are zero, this suggests that the k-th environment is non-intervened. The optimization process can be expressed as follows:

$$\min_{\mathbf{W}_{e_0}, \mathbf{W}_{e_k}} \sum_{k=1}^{n} \sum_{t=1}^{T} \sum_{\tau=1}^{l} ||\mathbf{X}_t - (\mathbf{W}_{e_0} + \mathbf{W}_{e_k})\mathbf{X}_{t-\tau}||_2^2 + \lambda\Omega(\mathbf{W}_{e_0}, \mathbf{W}_{e_k}), \tag{4}$$

where τ is the time lag. The final estimated Granger causal structure without interventions is represented by $\mathbf{W}_{e_0}$, and $\mathbf{W}_{e_k}$ denotes the underlying intervention structures after the k-th intervention. The implementation details of the regularization penalty term $\Omega(\mathbf{W}_{e_0}, \mathbf{W}_{e_k})$ will be discussed in the following section.

4.2 Non-linear Granger Causality with Interventions

Linear Granger causal models, with their simplicity and straightforwardness, provide a clear but often oversimplified view of relationships among variables. In practice, it is challenging to model the highly non-linear relationships among multiple variables from

time series data. In our proposed *IGC*, we assume that there exist functions $f_i : \mathbb{R}^{d\times T} \to \mathbb{R}$ and $g_i : \mathbb{R}^{d\times T} \to \mathbb{R}$ such that:

$$\mathbb{E}[x_{i,t+1}|\mathbf{PA}(x_{i,t+1})] = f_i(\mathbf{X}_{t:t-T}) + g_i(\mathbf{X}_{t:t-T}). \tag{5}$$

$f_i(\mathbf{x}_1, .., \mathbf{x}_d)$ does not depend on $\mathbf{x}_k \in \mathbb{R}^T$ if $\mathbf{x}_k \cap \mathbf{PA}(x_{i,t+1}) = \emptyset$; and $g_i(\mathbf{X}_{t:t-T}) = 0$ if $x_{i,t+1}$ is not intervened, under the assumption of no instantaneous effects [33]. Thus, our objective is to learn $f = (f_1, ..., f_d)$ and $g = (g_1, ..., g_d)$ such that the estimated Granger causal structure from f and interventional targets from g.

Let us first define $\overline{\mathbf{W}}_{e_k} = \mathbf{W}_{e_0} + \mathbf{W}_{e_k}$ and concentrate on a single variable x_i within a specific environment e_k. We define a set of parameters, which can be represented as: $\phi^i_{e_k}$, where $\phi^i_{e_k} = \{\overline{\mathbf{W}}^i_{:,1,e_k}, \ldots, \overline{\mathbf{W}}^i_{:,d,e_k}\}$ and $\phi_{e_k} = \{\phi^1_{e_k}, \ldots, \phi^d_{e_k}\}$. Then the overall objective becomes:

$$\min_{\phi_{e_k}} \sum_{i=1}^{d}\sum_{t=2}^{T} ||x^{e_k}_{i,t} - \mathbf{F}_i(x^{e_k}_{t-1:1}; \phi^i_{e_k})||_2^2 + \lambda \sum_{i=1}^{d}\sum_{j=1}^{d} ||\overline{\mathbf{W}}^i_{:,j,e_k}||_2, \tag{6}$$

where $\mathbf{F}_i(\cdot)$ is defined as:

$$\mathbf{F}_i(x^{e_k}; \phi^i_{e_k}) = f_i(x^{e_k}; \mathbf{W}^i_{e_0}) + g_i(x^{e_k}; \mathbf{W}^i_{e_k}), \tag{7}$$

and $\mathbf{F}_i(\cdot)$ generates the estimate $\hat{x}_i$ for the next timestep in e_k, time series j is *Granger non-causal* for time series i in the e_k if and only if $\overline{\mathbf{W}}^i_{:,j,e_k}$ is zero. With the above proposed objective and heterogeneous interventional time series data from n environments, we propose minimizing Equation (8) to prioritize the discovery of the Granger causal structure that remains consistent across all environments $\mathcal{E} = \{e_1, \ldots, e_n\}$.

$$\begin{aligned}\min_{\phi} &\sum_{k=1}^{n}\sum_{i=1}^{d}\sum_{t=2}^{T} ||x^{e_k}_{i,t} - \mathbf{F}_i(x^{e_k}_{t-1:1}; \phi^i_{e_k})||_2^2 \\ &+ \lambda \sum_{i=1}^{d}\sum_{j=1}^{d} ||(\overline{\mathbf{W}}^i_{:,j,e_1}, ..., \overline{\mathbf{W}}^i_{:,j,e_n})||_2,\end{aligned} \tag{8}$$

where $\phi = \{\phi_{e_1}, \ldots, \phi_{e_n}\}$ represents the collection of parameters across all n environments.

To learn the unknown interventional targets while maintaining consistency in the Granger causal structure, the overall penalized objective becomes:

$$\begin{aligned}\min_{\phi} &\sum_{k=1}^{n}\sum_{t=2}^{T}\sum_{i=1}^{d}\sum_{j=1}^{d} ||x^{e_k}_{i,t} - \mathbf{F}_{ij}(x^{e_k}_{j,t-1:1}; \overline{\mathbf{W}}^i_{:,j,e_k})||_2^2 \\ &+ (1-\alpha)\lambda \sum_{i=1}^{d}\sum_{j=1}^{d} ||(\mathbf{W}^i_{:,j,e_0}, \mathbf{W}^i_{:,j,e_1}, ..., \mathbf{W}^i_{:,j,e_n})||_2 \\ &+ \alpha\lambda \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{n} ||\mathbf{W}^i_{:,j,e_k}||_2,\end{aligned} \tag{9}$$

where $\alpha \in (0, 1)$ controls the tradeoff in sparsity across and within groups. After learning, time series j is *Granger non-causal* to i if $\mathbf{W}^i_{:,j,e_0}$ is zero across k distributions. Furthermore, there is no intervention from time series j to i in the k-th distribution if $\mathbf{W}^i_{:,j,e_k}$ is zero, which can be mathematically expressed as:

$$\mathbf{P}(x^{e_0}_{i,t}|x^{e_0}_{j,t-1:t-\tau}) = \mathbf{P}(x^{e_k}_{i,t}|x^{e_k}_{j,t-1:t-\tau}). \tag{10}$$

4.3 Model Architecture

In line with the concepts presented in Section 4.2, *IGC* utilizes historical time series data as its input and forecasts the data for the subsequent timestep as its output. The principal contribution of our study is the integration of heterogeneous interventional time series data, which aids in the identification of both the Granger causal structure and the interventional targets. The architecture of the model is illustrated in Figure 3.

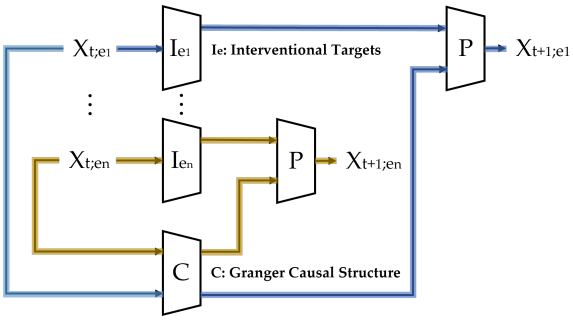

Figure 3: The information flow in various environments is represented by different colors. During the learning process, the prediction network (P) generates data for the next timestep. Information about unknown targets is contained within the intervention networks ($\mathbf{I}_e$), and the Granger causal structure is captured within the causal network (C).

Intervention Networks. Consider a set of time series data $\mathbf{X} = \{\mathbf{X}_{e_1}, \ldots, \mathbf{X}_{e_n}\}$ from n environments or distributions. For each specific environment e_k, there exists an intervention network $\mathbf{I}_{e_k} = \{\mathbf{I}^1_{e_k}, \ldots, \mathbf{I}^d_{e_k}\}$, where each function $\mathbf{I}^i_{e_k}$ is defined as:

$$\mathbf{I}^i_{e_k}(\mathbf{X}_{t;e_k}; \mathbf{W}^i_{e_k}) : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times h}. \tag{11}$$

In this context, $\mathbf{I}^i_{e_k}(\cdot)$ represents the intervention network for node i in e_k, $\mathbf{X}_{t;e_k} \in \mathbb{R}^{T\times d}$ is the historical multivariate time series data in e_k, and $\mathbf{W}^i_{e_k} \in \mathbb{R}^{d\times h}$ denotes the parameters of the intervention network $\mathbf{I}^i_{e_k}$.

Granger Causal Network. For the time series data set $\mathbf{X}$, a shared Granger causal network is applicable to all environments within $\mathbf{X}$. This network is defined as $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_d\}$, where each $\mathbf{C}_i$ is described by the function:

$$\mathbf{C}_i(\mathbf{X}_t; \mathbf{W}^i_{e_0}) : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times h}. \tag{12}$$

In this context, $\mathbf{X}_t$ represents the historical multivariate time series data from each environment within $\mathbf{X}$ presented in sequence, and $\mathbf{W}^i_{e_0} \in \mathbb{R}^{d\times h}$ are the parameters of the Granger causal network for node i.

Information Aggregator. After generating both the intervention information and the Granger causal information, we use a mechanism to aggregate them:

$$\mathbf{Z}^i_{t;e_k} = \mathbf{Agg}(\mathbf{I}^i_{e_k}(\mathbf{X}_{t;e_k}; \mathbf{W}^i_{e_k}), \mathbf{C}_i(\mathbf{X}_{t;e_k}; \mathbf{W}^i_{e_0})), \tag{13}$$

where $\mathbf{Agg}(\cdot)$ is an aggregation function and we have adopted summation in our experiments. We leave a learnable aggregation operator as a future research direction.

Prediction Network. The prediction network $\mathbf{P}$ is designed to forecast the i-th data point for the subsequent timestep:

$$\widehat{\mathbf{X}}^i_{t+1;e_k} = \mathbf{P}(\mathbf{Z}^i_{t;e_k}), \tag{14}$$

where $\widehat{\mathbf{X}}^i_{t+1;e_k} \in \mathbb{R}$ represents the predicted value at timestep $t+1$, while $\mathbf{Z}^i_{t;e_k} \in \mathbb{R}^{T\times h}$ denotes the aggregated embedding obtained from the previous step.

To enhance flexibility, components such as $\mathbf{I}_e(\cdot)$, $\mathbf{C}(\cdot)$, and $\mathbf{P}(\cdot)$ can be effectively modeled using a variety of neural network architectures, including MLP, LSTM, SENNs, and Transformer. In our experiments, we employed MLPs and trained the model with respect to Equation (9). As illustrated in Figure 3, information about unknown targets is obtained from the intervention networks $\mathbf{I}_e$, while the Granger causal structure is estimated from the causal network $\mathbf{C}$. The *IGC* operates under the Assumption 1.

Assumption 1. *(Causal Consistency). There exists a consistent causal structure and common parameters $\mathbf{W}_{e_0}$ across different environments. The dissimilarity between the parameters for one environment and the common parameters $\mathbf{W}_{e_0}$ lies within a range defined by a lower bound ϵ_l, and an upper bound ϵ_u. This range captures the extent of variation allowed between the common parameters and different environments. To avoid identical data across environments, the condition $\epsilon_l = 0$ indicates that there is no significant intervention. Mathematically it can be expressed as: $\epsilon_l \leq |\mathbf{W}_{e_k}| \leq \epsilon_u, \forall 1 \leq k \leq n, 0 \leq \epsilon_l \leq \epsilon_u$.*

4.4 Optimizing the Penalized Objective

To optimize the objective stated in Equation (9) for the proposed *IGC* method, we use proximal gradient descent [32], which is particularly beneficial for our purposes as it results exact zeros in the columns of input parameters, an essential aspect for interpreting Granger non-causality and intervention within our framework. The proximal operator is the group-wise soft-thresholding operator. Detailed updates of the proximal gradient descent are included in the Appendix A.1. The proximal steps on the input weights for the penalty in Equation (9) is shown in Algorithm 1, where **Soft**$(\cdot)$ is a group soft-thresholding operator on the input weights [32].

Algorithm 1 Proximal steps for the penalty in Equation (9)

procedure Input($\alpha > 0, \lambda > 0, (\mathbf{W}^i_{:,j,e_0}, ..., \mathbf{W}^i_{:,j,e_k})$)
  for $k = 1$ to n **do**
    $\mathbf{W}^i_{:,j,e_k} = \mathbf{Soft}_{\alpha\lambda}(\mathbf{W}^i_{:,j,e_k})$
  end for
  $(\mathbf{W}^i_{:,j,e_0}, ..., \mathbf{W}^i_{:,j,e_k}) = \mathbf{Soft}_{(1-\alpha)\lambda}((\mathbf{W}^i_{:,j,e_0}, ..., \mathbf{W}^i_{:,j,e_k}))$
  return $(\mathbf{W}^i_{:,j,e_0}, ..., \mathbf{W}^i_{:,j,e_k})$
end procedure

5 IDENTIFIABILITY

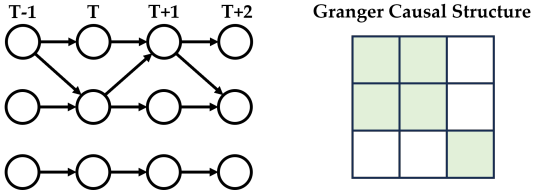

Figure 4: The complex interactions in time series data (left) lead to a Granger causal structure (right) that is not a strict DAG.

The identifiability of Granger causal structure for observational time series data has been established, where the parameters $\mathbf{W}$ can be identified from standard results in vector autoregressive (VAR) models [31]. For linear time series interventional data, the identifiability results have been studied in [3]. Specifically, the model is identifiable if each variable is influenced by a unique set of intervened variables. In the context of non-linear interventional time series data with known interventional targets, [12] expanded upon the $\mathcal{I}$-Markov Equivalence Class to $(\mathcal{I}, \mathcal{D})$-Markov Equivalence Class for graphs within a subset of DAGs rather than all DAGs. However, addressing the challenge of identifying Granger causal structure in non-linear time series data with unknown interventional targets remains a significant and unresolved area of research. To address this issue, our initial step is to establish the negative score function for a DAG $\mathcal{G}$:

$$-\mathcal{S}_{\mathcal{I}}(\mathcal{G}) := \mathbb{E}_{\mathbf{X}|\mathbf{P}_{e_k},\mathcal{G}^*}[\mathcal{L}_{reg}(\mathbf{X})], \tag{15}$$

where $\mathcal{L}_{reg}$ denotes the regularized loss minimized in Equation (9), with the loss $||\cdot||_2^2$ being negative log-liklihood, over the time series data $\mathbf{X}$, which is generated from the ground truth $\mathcal{G}^*$, under the interventional distribution $\mathbf{P}_{e_k}$ for each $\mathbf{I}_k \in \mathcal{I}$. Based on these definitions, the following theorem holds:

Theorem 5.1. *Let $\hat{\mathcal{G}} \in \mathcal{D}$ be a DAG and $\hat{\mathcal{I}}$ be an interventional family, which $(\hat{\mathcal{G}}, \hat{\mathcal{I}}) \in \arg\max_{\mathcal{G},\mathcal{I}} \mathcal{S}(\mathcal{G}, \mathcal{I})$. Under the assumption that the density models have sufficient capacity to represent the ground truth distribution, that $\mathcal{I}^*$-faithfulness holds, that the density models are strictly positive, that the ground truth densities $\mathbf{P}_{e_k}$ have differentiable entropy. For $\lambda_{\mathcal{G}}, \lambda_{\mathcal{I}} > 0$ in Equation (9) small enough, $\hat{\mathcal{G}}$ is $(\mathcal{I}^*, \mathcal{D})$-Markov equivalent to $\mathcal{G}^*$ and $\hat{\mathcal{I}} = \mathcal{I}^*$.*

The Granger causal strcture we've learned is not a strict DAG due to the intricate nature of time series data, as shown in Figure 4. However, rather than focusing directly on the Granger causal structure, our approach centers on the complex interactions within the time series data. Particularly, given the forward-in-time property, the unrolled temporally extended graph is a DAG $\in \mathbb{R}^{d\times T}$ and does not include any cyclic subgraphs, thus we omit the DAG constraint [50] in our theoretical analysis. Establishing the identifiability of this DAG also allows us to identify the Granger causal structure. The *IGC* methodology, characterized by its time-windowed approach, offers flexibility for detecting the causal relationship between any two variables $(x_{i,t}, x_{j,t'})$ in this DAG with a given time lag $p = t' - t$. If we set $\mathcal{D}$ to be the subset $\mathcal{D}_s$ of DAGs which correspond to stationary dynamics with constant-in-time conditional distributions (For detailed information and the proof of Theorem 5.1, please refer to the Appendix A.2), Theorem 5.1 can be restated as follows:

Corollary 5.2. *Let $\hat{\mathcal{G}} \in \mathcal{D}_s$ be a DAG and $\hat{\mathcal{I}}$ be an interventional family. Given the same assumptions as Theorem 5.1, and for $\lambda_{\mathcal{G}}, \lambda_{\mathcal{I}}$ in Equation (9) small enough, $\hat{\mathcal{G}}$ is $(\mathcal{I}^*, \mathcal{D}_s)$-Markov equivalent to $\mathcal{G}^*$ and $\hat{\mathcal{I}} = \mathcal{I}^*$.*

The Theorem 5.1 extends prior work [2, 12] by showing that, under appropriate assumptions, maximizing $\mathcal{S}(\mathcal{G}, \mathcal{I})$ with respect $\mathcal{G}$ and $\mathcal{I}$ recovers both the $(\mathcal{I}^*, \mathcal{D})$-Markov equivalent class of $\mathcal{G}^*$ and the ground truth interventional family $\mathcal{I}^*$.

Dataset	Metrics	VAR	PCMCI	NGC	eSRU	DyNoTears	GVAR	CUTS	**IGC**
Linear (n=5)	Acc	0.640(±0.080)	0.800(±0.040)	0.920(±0.000)	0.960(±0.000)	0.800(±0.040)	0.960(±0.000)	0.920(±0.000)	**1.000(±0.000)**
	AUROC	0.650(±0.017)	0.770(±0.012)	0.925(±0.011)	0.967(±0.008)	0.740(±0.005)	0.985(±0.015)	0.933(±0.005)	**1.000(±0.000)**
	F1	0.609(±0.008)	0.667(±0.017)	0.909(±0.000)	0.952(±0.000)	0.725(±0.024)	0.949(±0.000)	0.911(±0.000)	**1.000(±0.000)**
	SHD	9(±2)	5(±1)	2(±0)	1(±0)	5(±1)	1(±0)	2(±0)	**0(±0)**
Linear (n=10)	Acc	0.560(±0.030)	0.610(±0.030)	0.820(±0.040)	0.850(±0.050)	0.650(±0.020)	**0.930(±0.010)**	0.880(±0.020)	**0.930(±0.010)**
	AUROC	0.562(±0.024)	0.710(±0.012)	0.848(±0.010)	0.812(±0.008)	0.524(±0.006)	0.980(±0.013)	0.865(±0.042)	**0.989(±0.018)**
	F1	0.551(±0.029)	0.456(±0.048)	0.847(±0.014)	0.869(±0.022)	0.596(±0.032)	0.912(±0.014)	0.872(±0.012)	**0.928(±0.017)**
	SHD	44(±3)	39(±3)	18(±4)	15(±5)	35(±2)	7(±1)	12(±2)	7(±1)
Linear (n=20)	Acc	0.518(±0.030)	0.555(±0.030)	0.815(±0.030)	0.730(±0.020)	0.565(±0.023)	0.783(±0.040)	0.838(±0.020)	**0.955(±0.005)**
	AUROC	0.538(±0.035)	0.545(±0.035)	0.822(±0.011)	0.723(±0.035)	0.511(±0.005)	0.854(±0.019)	0.832(±0.017)	**0.973(±0.006)**
	F1	0.671(±0.012)	0.351(±0.052)	0.812(±0.000)	0.772(±0.012)	0.322(±0.046)	0.800(±0.038)	0.816(±0.011)	**0.955(±0.002)**
	SHD	193(±6)	178(±12)	74(±12)	108(±6)	174(±9)	87(±16)	65(±8)	**18(±2)**
Dataset	**Metrics**	VAR	PCMCI	NGC	eSRU	DyNoTears	GVAR	CUTS	**IGC**
Non-linear (n=5)	Acc	0.458(±0.080)	0.560(±0.040)	0.960(±0.000)	0.760(±0.040)	0.800(±0.080)	0.920(±0.040)	0.920(±0.040)	**1.000(±0.000)**
	AUROC	0.517(±0.035)	0.567(±0.009)	0.967(±0.008)	0.767(±0.018)	0.740(±0.005)	0.912(±0.019)	0.935(±0.015)	**1.000(±0.000)**
	F1	0.563(±0.013)	0.522(±0.012)	0.952(±0.000)	0.727(±0.006)	0.725(±0.054)	0.920(±0.020)	0.915(±0.016)	**1.000(±0.000)**
	SHD	14(±2)	11(±1)	1(±0)	6(±1)	5(±2)	2(±1)	2(±1)	**0(±0)**
Non-linear (n=10)	Acc	0.520(±0.020)	0.580(±0.020)	0.880(±0.020)	0.710(±0.030)	0.620(±0.030)	0.920(±0.010)	0.860(±0.030)	**0.930(±0.020)**
	AUROC	0.512(±0.004)	0.626(±0.015)	0.892(±0.009)	0.709(±0.038)	0.548(±0.008)	0.901(±0.020)	0.859(±0.031)	**0.959(±0.005)**
	F1	0.658(±0.029)	0.600(±0.020)	0.893(±0.011)	0.721(±0.012)	0.498(±0.042)	0.913(±0.016)	0.834(±0.009)	**0.942(±0.011)**
	SHD	48(±2)	42(±2)	12(±2)	29(±3)	38(±3)	9(±1)	14(±3)	7(±2)
Non-linear (n=20)	Acc	0.508(±0.008)	0.545(±0.025)	0.795(±0.018)	0.647(±0.013)	0.543(±0.020)	0.825(±0.048)	0.805(±0.020)	**0.943(±0.008)**
	AUROC	0.515(±0.010)	0.548(±0.020)	0.800(±0.014)	0.641(±0.014)	0.587(±0.008)	0.882(±0.016)	0.820(±0.035)	**0.950(±0.015)**
	F1	0.659(±0.008)	0.461(±0.022)	0.793(±0.020)	0.714(±0.003)	0.435(±0.017)	0.821(±0.027)	0.811(±0.004)	**0.944(±0.006)**
	SHD	197(±3)	182(±10)	82(±7)	141(±5)	183(±8)	70(±19)	78(±8)	**23(±3)**

Table 1: Comparative results (mean ± std.) for synthetic interventional datasets.

6 EXPERIMENTS

We evaluate our proposed *IGC*[1] for inferring Granger causal structure and compare them with various state-of-the-art (SOTA) baselines across several interventional time series datasets, demonstrating the superior performance of our proposed *IGC* method. The competing SOTA methods for learning Granger causal structure that we benchmarked are listed as follows: 1) **VAR** (Vector AutoRegressive) [15, 16] is a linear model used in Granger causality test. **PCMCI** [38] integrates conditional independence tests with optimized conditioning sets for inferring causal structure. **NGC** [39] includes the component-wise MLP and the component-wise LSTM, featuring sparse input weight layers, is proposed as an effective approach for inferring non-linear Granger causality. **eSRU** [21] (economy Statistical Recurrent Units) are a specialized form of recurrent neural networks (RNNs) tailored to identify the network structure of non-linear Granger causal relationships. **DyNoTears** [31] is a score-based method with continuous optimization for learning causal structure. **GVAR** [28] model integrates SENNs with traditional vector autoregression for Granger causal inference. **CUTS** [4, 5] is a neural Granger causal discovery algorithm for imputed and high dimensional data.

For evaluation purposes, we utilize the following metrics: **Accuracy** refers to the rate at which a model correctly predicts the presence or absence of edges in the ground-truth graph. **AUROC** (Area Under the Receiver Operating Characteristic) curve is represented by the area under a curve plotting the true positive rate against the false positive rate at various thresholds. **AUPRC** (Area Under the Precision-Recall Curve) focuses on the relationship between precision and recall across different thresholds. The **F1 Score** represents the harmonic mean of precision and recall, with precision being the proportion of correctly detected edges relative to all edges predicted by the model. **SHD** (Structural Hamming Distance) denotes the count of incorrectly predicted edge states. **Recall** measures the fraction of edges in the ground-truth graph that are accurately identified by the model.

[1]https://github.com/Tamuzzy/IGC

6.1 Granger Causal Structure Learning

Firstly, we follow the functional causal model [18] detailed in Equation (16) to generate synthetic interventional time series data:

$$X_t^i = \sum f_i(X_j) + \epsilon_{i,t}, \tag{16}$$

where $X_j \in \mathbf{PA}(X_{i,t})$ and the function f_i can be selected from a set of functions which includes linear, cubic, tanh, and sinc functions, as well as their mixtures. The noise term $\epsilon_{i,t}$ is generated from either a uniform distribution $\mathcal{U}(-0.5, 0.5)$ or a standard normal distribution $\mathcal{N}(0, 1)$.

Linear Synthetic Interventional Time Series Data: In the linear setting, we generate the time series data by following these steps:

- We constructed the Granger causal graph $\mathcal{G}$ by employing two tunable parameters: n (number of nodes) and p (probability of edge creation). In our experiments, we set n to be 5, 10, and 20, and $p = 0.4$, we sample the weights uniformly at random from $\mathcal{U}([-0.6, -0.4] \cup [0.4, 0.6])$.
- We generate the data with first autoregessive order, where data only depends on the previous time step. We generate 5, 10, and 20 sequences with 500 time steps with standard Gaussian noises.
- To generate the interventional time series data in e_k, we adopt an imperfect setting where the weights from $\mathbf{PA}(X^i_{t;e_k})$ to $X^i_{t;e_k}$ are altered at a specific timestep $t = 200$ by adding a random number within the range $\mathcal{U}([-0.15, 0) \cup (0, 0.15])$.

Table 1 illustrates that *IGC* achieves the best performance, and performs slightly better at $n = 20$ than at $n = 10$, which suggests that the distinction does not significantly weaken our method.

Non-linear Synthetic Interventional Time Series Data: To evaluate the efficiency of *IGC* in the context of non-linear synthetic interventional time series data, we follow the same generation procedure as that used in the linear setting. However, instead of employing a linear function, we modify the underlying generation function f_i in Equation (16) to employ a 2-layer fully connected neural network with the Leaky ReLU activation and 0.1 negative non-linearity. The network weights are sampled uniformly from $\mathcal{U}([-0.6, -0.4] \cup [0.4, 0.6])$. To implement imperfect interventions, we add a random vector, drawn from $\mathcal{N}(0, 1)$ to the network's second layer.

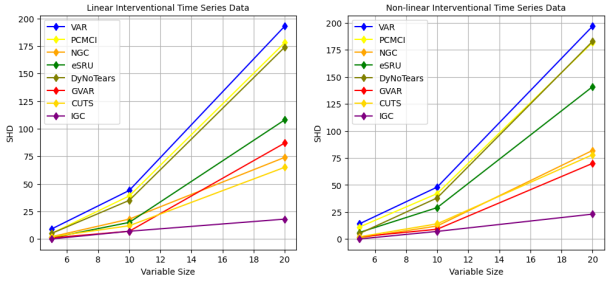

Figure 5: SHD results for Linear (left) and Non-linear (right) Synthetic Interventional Time Series Data.

As illustrated in Table 1, we compare the performance of *IGC* against other methods for $n = \{5, 10, 20\}$. The results confirm that our method consistently outperforms the others, even when the variable size is large. Figure 5 presents the comparison of the results for learning Granger causal structure with varying numbers of variables in both linear and non-linear settings. From the results, we observe that the introduction of interventional data disrupts the stationary assumption underlying these models, leading to poor performance in inferring the Granger causal structure. In contrast, our model exhibits enhanced capability in managing the non-stationarity induced by interventions, thereby achieving more accurate inference of the Granger causal structure.

Lorenz-96 Model				
Metric/Methods	Acc	AUPRC	AUROC	SHD
VAR	0.765(±0.008)	0.464(±0.046)	0.745(±0.047)	94(±3)
PCMCI	0.720(±0.020)	0.724(±0.007)	0.788(±0.033)	112(±8)
NGC	0.653(±0.028)	0.956(±0.016)	0.979(±0.016)	139(±11)
eSRU	0.823(±0.010)	0.834(±0.033)	0.934(±0.021)	70(±4)
DyNoTears	0.785(±0.013)	0.779(±0.035)	0.811(±0.015)	86(±5)
GVAR	0.845(±0.010)	0.916(±0.024)	0.970(±0.009)	62(±4)
CUTS	0.755(±0.023)	0.785(±0.015)	0.876(±0.017)	98(±9)
IGC	**0.925(±0.008)**	**0.979(±0.003)**	**0.985(±0.002)**	**30(±2)**

Table 2: Comparative results (mean±std.) for Lorenz-96.

Lorenz-96 Model: The Lorenz 96 model, a standard benchmark for Granger causal inference techniques [25], is a continuous-time dynamic system with m variables, defined by non-linear differential equations:

$$\frac{dx^i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \tag{17}$$

where $x_0 := x_m$, $x_{-1} := x_{m-1}$, and $x_{m+1} := x_1$; and F is a forcing constant that, in combination with m, controls the non-linearity of the system [20, 39]. We numerically simulate $m = 20$ variables and $T = 500$ observations under $F = 40$. This choice is predicated on the understanding that a higher number of variables coupled with a higher non-linearity ($F = 40$) presents a more challenging inference problem. While adhering to the experimental setup of [28], our study introduces a more challenging setting. We manipulated the data with $m = 20$ variables and $T = 500$ observations under $F = 40$ by altering the value of F to 50 for samples when $t > 250$, thus introducing an intervention in the dataset to simulate real-world complexities. From Table 2, we observe that our proposed *IGC* achieves competitive performance even in more complex situations.

Tennessee Eastman Dataset									
Metric/Methods	Acc	Recall	F1	SHD	**Metric/Methods**	Acc	Recall	F1	SHD
CORL	0.838	0.043	0.071	176	NoTears-MLP	0.925	0.036	0.046	82
DirectLiNGAM	0.918	0.046	0.061	89	PCMCI	0.882	0.094	0.044	129
FCI	0.966	0.167	0.091	37	DyNoTears	0.928	0.094	0.071	78
GES	0.903	0.040	0.060	106	eSRU	0.936	0.054	0.068	70
GEOLEM	0.890	0.031	0.046	120	GVAR	0.928	0.188	0.133	78
ICALiNGAM	0.908	0.079	0.116	100	NGC	0.852	0.089	**0.148**	161
MCSL	0.951	0.080	0.070	53	CUTS	0.922	0.094	0.066	85
NoTears	0.968	0	0	35	**IGC**	**0.968**	**0.286**	0.103	**35**

Table 3: Comparative results for TEP Dataset.

Tennessee Eastman Process (TEP): The Tennessee Eastman Process (TEP) [7], serves as a widely recognized benchmark in chemical engineering research. This simulator is particularly valuable for studies in anomaly detection and root cause analysis, due to its capability to replicate process faults and the comprehensive description it offers of the entire production process. The TEP includes five principal units: a two-phase reactor, a condenser, a recycle compressor, a liquid-vapor separator, and a product stripper, involving 41 measured and 12 manipulated variables. The observational dataset is devoid of anomalies and comprises 500 observations. Within the TEP, there are 21 predefined faults, resulting in 21 distinct test datasets. Each dataset contains 960 observations, recorded at 3-minute intervals. The initial 160 observations in each dataset are anomaly-free. Starting from observation 161 and continuing to the end of the dataset, one of the 21 faults is introduced, marking a transition to conditions where the system's behavior deviates from the norm. In our research, we have utilized 22 measured variables and 11 manipulated variables. We have employed various causal structure learning methods on the observational data and integrated our propsoed *IGC* approach on both observational data and interventional data with anomalies. The results, summarized in Table 3, demonstrate that our method consistently outperforms several other techniques, reinforcing its efficacy in handling interventional time series data. The higher Recall and F1-Score of our proposed method compared to NoTears can be attributed to a greater number of True Positives (TP) since the TEP dataset with a quite sparse adjacency matrix $\in \mathbb{R}^{33\times33}$, where only 66 elements being 1 (including 33 diagonal elements). Our method, which is based on Granger

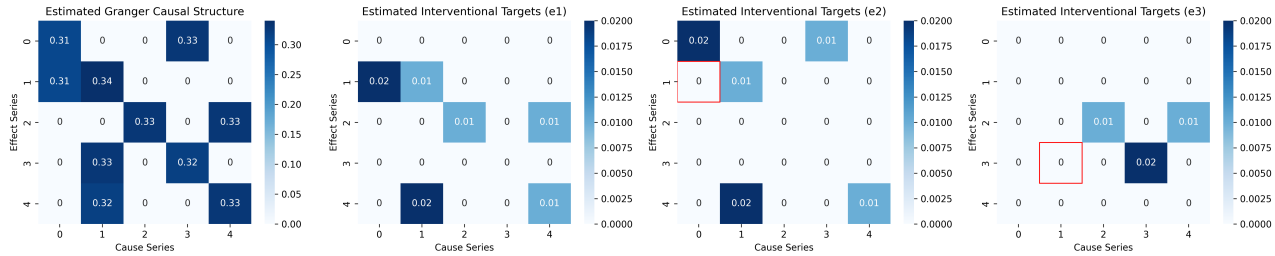

Figure 6: The estimated Granger causal structure and the estimated unknown interventional targets across three different environments, we have highlighted the discrepancies in the results using red blocks to indicate areas of disagreement.

causality, leverages historical temporal information effectively. In contrast, NoTears is not specifically designed for time series data, leading to its less efficient capture of this crucial information.

6.2 Interventional Family Recovery

So far, our focus has been on inferring Granger causal structure, without addressing the issue of interventional family recovery, which is crucial for a deeper understanding of various time series analysis tasks. Although there have been some experiments targeting known interventions, to the best of our knowledge, this study is the first to delve into the recovery of unknown interventional targets from interventional time series data. This approach not only enhances our understanding of the underlying processes but also clarifies the interaction between the Granger causal structure learning and the recovery of unknown targets. To bridge the identified gap, we assessed the model's capability in accurately identifying unknown interventional targets within synthetic datasets. We first formulate the problem as follows:

Problem 1. Consider a Granger causal graph $\mathcal{G} \in \mathbb{R}^{d\times d}$, with the assumption that the time series generation follows Equation (16). We introduce *interventional family*, denoted as $\mathcal{I} := \{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$, where each $\mathbf{I}_k \in \mathbb{R}^{d\times d}$. This implies that, after t time steps, some specific edges in the Granger causal graph $\mathcal{G}$ are chosen as interventional targets in k-th intervention. The problem is to recover these interventional targets based solely on the observed intervened or non-intervened time series data from several environments, without access to the knowledge of non-intervened time series data.

Figure 6 illustrated that the Granger causal structure, generated from the causal network, is stable across multiple environments, despite changes in causal strength following the interventions. Regarding the intervention networks, the results highlight the effects of interventions. We used red blocks to indicate areas of disagreement in Figure 6 and we attribute these discrepancies to instances where the intervention strength was insufficient or below certain thresholds. We also evaluated the task of recovering interventional targets on synthetic interventional time series data from several distinct environments, as illustrated in Figure 7. We found that our method prioritizes environments experiencing high-intensity interventions, potentially overlooking those with milder interventions. Thus, it is important to set thresholds based on the specific application to determine whether the environment is being intervened.

It is also worth noting that the number of interventions should be less than a threshold, which can be described as: $|\mathcal{I}| \leq n$.

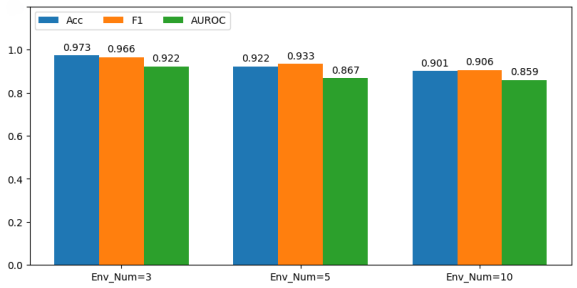

Figure 7: Evaluation of the interventional family recovery.

7 CONCLUSION

In this study, we have investigated the Granger causal structure learning task, incorporating heterogeneous interventional time series data. To address the issue of identifying Granger non-causality in interventional time series data with unknown targets, we have introduced a novel condition that ensures the recovery of these unknown targets and the accurate identification of the true causal structure within the $(\mathcal{I}, \mathcal{D})$-Markov Equivalence Class. We solved the identifiability issues for accurately determining causal relationships in Granger causality. Our theoretical analysis is supported by empirical results, demonstrating that our proposed Interventional Granger Causal (*IGC*) structure learning method outperforms existing methodologies in both synthetic and real-world datasets, even in the absence of interventional target information. Potential avenues for future research include applying our method to a broader spectrum of time series applications, which includes detecting anomalies and root cause analysis within time series data.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. National Science Foundation (NSF) grants SHF-2215573, and by the U.S. Department of Engergy (DOE) Office of Science, Advanced Scientific Computing Research (ASCR) under Awards B&R# KJ0403010/FWP#CC132 and FWP#CC138. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

REFERENCES

[1] Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal causal modeling with graphical granger methods. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 66–75.

[2] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. 2020. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems* 33 (2020), 21865–21877.

[3] Chen Chen, Min Ren, Min Zhang, and Dabao Zhang. 2018. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research* 19, 1 (2018), 40–73.

[4] Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. 2024. CUTS+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11525–11533.

[5] Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. 2023. Cuts: Neural causal discovery from irregular time-series data. *arXiv preprint arXiv:2302.07458* (2023).

[6] Yunfei Chu, Xiaowei Wang, Jianxin Ma, Kunyang Jia, Jingren Zhou, and Hongxia Yang. 2020. Inductive granger causal modeling for multivariate time series. In *IEEE International Conference on Data Mining*. IEEE, 972–977.

[7] James J Downs and Ernest F Vogel. 1993. A plant-wide industrial process control problem. *Computers & Chemical Engineering* 17, 3 (1993), 245–255.

[8] Daniel Eaton and Kevin Murphy. 2007. Exact Bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*. PMLR, 107–114.

[9] Frederick Eberhardt. 2008. Almost optimal intervention sets for causal discovery. *Conference on Uncertainty in Artificial Intelligence*.

[10] Frederick Eberhardt, Clark Glymour, and Richard Scheines. 2005. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *Conference on Uncertainty in Artificial Intelligence*.

[11] Frederick Eberhardt and Richard Scheines. 2007. Interventions and causal inference. *Philosophy of Science* 74, 5 (2007), 981–995.

[12] Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, and Yue Yu. 2022. IDYNO: Learning nonparametric DAGs from interventional dynamic data. In *International Conference on Machine Learning*. PMLR, 6988–7001.

[13] AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. 2018. Multi-domain causal structure learning in linear systems. *Advances in Neural Information Processing Systems* 31 (2018).

[14] Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, Jingping Bi, Lun Du, and Jin Wang. 2023. Causal Discovery from Temporal Data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5803–5804.

[15] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[16] Clive WJ Granger. 1980. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* 2 (1980), 329–352.

[17] Alain Hauser and Peter Bühlmann. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 13, 1 (2012), 2409–2464.

[18] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. 2020. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research* 21, 1 (2020), 3482–3534.

[19] Iris AM Huijben, Arthur Andreas Nijdam, Sebastiaan Overeem, Merel M Van Gilst, and Ruud Van Sloun. 2023. Som-cpc: Unsupervised contrastive learning with self-organizing maps for structured representations of high-rate time series. In *International Conference on Machine Learning*. PMLR, 14132–14152.

[20] Alireza Karimi and Mark R Paul. 2010. Extensive chaos in the Lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20, 4 (2010).

[21] Saurabh Khanna and Vincent YF Tan. 2019. Economy Statistical Recurrent Units For Inferring Nonlinear Granger Causality. In *International Conference on Learning Representations*.

[22] Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. 2004. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 322–331.

[23] Hongming Li, Shujian Yu, and Jose Principe. 2023. Causal recurrent variational autoencoder for medical time series generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[24] Chenxi Liu and Kun Kuang. 2023. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*. PMLR, 21756–21777.

[25] Edward N Lorenz. 1996. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, Vol. 1. Reading.

[26] Qianli Ma, Sen Li, and Garrison W Cottrell. 2020. Adversarial joint-learning recurrent neural network for incomplete time series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 1765–1776.

[27] Qianli Ma, Jiawei Zheng, Sen Li, and Gary W Cottrell. 2019. Learning representations for time series clustering. *Advances in Neural Information Processing Systems* 32 (2019).

[28] Ričards Marcinkevičs and Julia E Vogt. 2020. Interpretable Models for Granger Causality Using Self-explaining Neural Networks. In *International Conference on Learning Representations*.

[29] Xinyi Ni and Lifeng Lai. 2022. Policy gradient based entropic-var optimization in risk-sensitive reinforcement learning. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 1–6.

[30] Xinyi Ni and Lifeng Lai. 2022. Risk-sensitive reinforcement learning via Entropic-VaR optimization. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 953–959.

[31] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1595–1605.

[32] Neal Parikh, Stephen Boyd, et al. 2014. Proximal algorithms. *Foundations and Trends in Optimization* 1, 3 (2014), 127–239.

[33] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems* 26 (2013).

[34] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[35] Huida Qiu, Yan Liu, Niranjan A Subrahmanya, and Weichang Li. 2012. Granger causality for time-series anomaly detection. In *International Conference on Data Mining*. IEEE, 1074–1079.

[36] Shaogang Ren and Ping Li. 2022. Flow-based perturbation for cause-effect inference. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1706–1715.

[37] Shaogang Ren, Haiyan Yin, Mingming Sun, and Ping Li. 2021. Causal discovery with flow-based conditional density estimation. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1300–1305.

[38] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* 5, 11 (2019), eaau4996.

[39] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. 2021. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4267–4279.

[40] Thomas S Verma and Judea Pearl. 2022. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The works of Judea Pearl*. 221–236.

[41] E Vijay, Arindam Jati, Nam Nguyen, Gift Sinthong, and Jayant Kalagnanam. 2023. TSMixer: lightweight MLP-mixer model for multivariate time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[42] Song Wei, Yao Xie, Christopher S Josef, and Rishikesan Kamaleswaran. 2023. Granger causal chain discovery for sepsis-associated derangements via continuous-time Hawkes processes. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2536–2546.

[43] Chenxiao Xu, Hao Huang, and Shinjae Yoo. 2019. Scalable causal graph learning through a deep neural network. In *ACM International Conference on Information and Knowledge Management*. 1853–1862.

[44] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. 2021. Voice2series: Reprogramming acoustic models for time series classification. In *International Conference on Machine Learning*. PMLR, 11808–11819.

[45] Karren Yang, Abigail Katcoff, and Caroline Uhler. 2018. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*. PMLR, 5541–5550.

[46] Raneen Younis, Zahra Ahmadi, Abdul Hakmeh, and Marco Fisichella. 2023. Flames2graph: An interpretable federated multivariate time series classification framework. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3140–3150.

[47] Ziyi Zhang, Diya Li, Zhenlei Song, Nick Duffield, and Zhe Zhang. 2023. Location-Aware Social Network Recommendation via Temporal Graph Networks. In *ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*. 58–61.

[48] Ziyi Zhang, Diya Li, Zhe Zhang, and Nicholas Duffield. 2021. A time-series clustering algorithm for analyzing the changes of mobility pattern caused by COVID-19. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop On Animal Movement Ecology And Human Mobility*. 13–17.

[49] Ziyi Zhang, Shaogang Ren, Xiaoning Qian, and Nick Duffield. 2024. Towards Invariant Time Series Forecasting in Smart Cities. In *Companion Proceedings of the ACM on Web Conference 2024*. 1344–1350.

[50] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems* 31 (2018).

[51] Alex M Zimmer, Yihang K Pan, Theanuga Chandrapalan, Raymond WM Kwong, and Steve F Perry. 2019. Loss-of-function approaches in comparative physiology: is there a future for knockdown experiments in the era of genome editing? *Journal of Experimental Biology* 222, 7 (2019), jeb175737.

A APPENDIX

A.1 Proximal Gradient Descent Updates

As for Equation (9), we establish the following definitions:

$$g(\phi) := \sum_{k=1}^{n}\sum_{t=2}^{T}\sum_{i=1}^{d}\sum_{j=1}^{d} ||x_{i,t}^{e_k} - \mathbf{F}_{ij}\left(x_{j,t-1:1}^{e_k}; (\mathbf{W}_{:,j,e_0}^{i} + \mathbf{W}_{:,j,e_k}^{i})\right)||_2^2. \tag{18}$$

$$\begin{aligned} h(\phi) :&= (1-\alpha)\lambda \sum_{i=1}^{d}\sum_{j=1}^{d} ||(\mathbf{W}_{:,j,e_0}^{i}, \mathbf{W}_{:,j,e_1}^{i}, ..., \mathbf{W}_{:,j,e_k}^{i})||_2 \\ &+ \alpha\lambda \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{n} ||\mathbf{W}_{:,j,e_k}^{i}||_2. \end{aligned} \tag{19}$$

A proximal mapping for the function $h(\phi)$ can be defined as follows:

$$\begin{aligned} \mathbf{prox}_h(\mathbf{u}) = \underset{\mathbf{z}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{z}-\mathbf{u}||_2^2 + (1-\alpha)\lambda||\mathbf{z}||_2 \\ + \alpha\lambda \sum_{i=1}^{n} ||z_i||_2. \end{aligned} \tag{20}$$

For $k = 0, 1, \ldots, n$, the updating steps at the m-th iteration are represented as:

$$\mathbf{W}_{:,j,e_k}^{i(m)} = \mathbf{prox}_{h,t_k}(\mathbf{W}_{:,j,e_k}^{i(m-1)} - t_k \nabla_{\mathbf{W}_{:,j,e_k}^{i}} g(\phi^{(m-1)})). \tag{21}$$

Thus, $\mathbf{u} = \{u_0, u_1, \ldots, u_n\}$ is a vector, and it is defined as:

$$u_k = \mathbf{W}_{:,j,e_k}^{i(m-1)} - t_k \nabla_{\mathbf{W}_{:,j,e_k}^{i}} g(\phi^{(m-1)}), \tag{22}$$

and

$$\mathbf{W}_{:,j,e_k}^{i(m)} = \mathbf{prox}_h(u_k).$$

First, let's examine the scenario when $\mathbf{z} = 0$. According to the Karush-Kuhn-Tucker (KKT) conditions, we obtain the following:

$$0 \in \begin{bmatrix} z_0 \\ z_1 \\ \cdots \\ z_n \end{bmatrix} - \begin{bmatrix} u_0 \\ u_1 \\ \cdots \\ u_n \end{bmatrix} + (1-\alpha)\lambda \begin{bmatrix} \frac{z_0}{||\mathbf{z}||_2} \\ \frac{z_1}{||\mathbf{z}||_2} \\ \cdots \\ \frac{z_n}{||\mathbf{z}||_2} \end{bmatrix} + \alpha\lambda \begin{bmatrix} 0 \\ \frac{z_1}{||z_1||_2} \\ \cdots \\ \frac{z_n}{||z_n||_2} \end{bmatrix}. \tag{23}$$

One could set $\mathbf{z} = 0$, while Equation (24) holds:

$$\begin{bmatrix} u_0 \\ u_1 \\ \cdots \\ u_n \end{bmatrix} - \alpha\lambda \begin{bmatrix} 0 \\ \frac{z_1}{||z_1||_2} \\ \cdots \\ \frac{z_n}{||z_n||_2} \end{bmatrix} = (1-\alpha)\lambda \begin{bmatrix} \frac{z_0}{||\mathbf{z}||_2} \\ \frac{z_1}{||\mathbf{z}||_2} \\ \cdots \\ \frac{z_n}{||\mathbf{z}||_2} \end{bmatrix}. \tag{24}$$

Identifying edge cases for $\mathbf{u}$ is straightforward as it involves an element-wise comparison between $\mathbf{u}$ and $\mathbf{z}$. Additionally, it is worth noting that $||\mathbf{z}||_2 \leq 1$, leading to the following considerations:

$$\mathbf{z} = 0 \Leftrightarrow ||\mathbf{u} - \alpha\lambda \begin{bmatrix} 0 \\ \frac{u_1}{||u_1||_2} \\ \cdots \\ \frac{u_n}{||u_n||_2} \end{bmatrix}||_2 \leq (1-\alpha)\lambda. \tag{25}$$

In the case $\mathbf{z} \neq 0$, Equation (23) suggests:

$$\begin{bmatrix} u_0 \\ u_1 \\ \cdots \\ u_n \end{bmatrix} - \alpha\lambda \begin{bmatrix} 0 \\ \frac{z_1}{||z_1||_2} \\ \cdots \\ \frac{z_n}{||z_n||_2} \end{bmatrix} = (1-\alpha)\lambda \begin{bmatrix} \frac{z_0}{||\mathbf{z}||_2} \\ \frac{z_1}{||\mathbf{z}||_2} \\ \cdots \\ \frac{z_n}{||\mathbf{z}||_2} \end{bmatrix} + \begin{bmatrix} z_0 \\ z_1 \\ \cdots \\ z_n \end{bmatrix}. \tag{26}$$

When considering elements in $\mathbf{z}$ that are non-zero, their sign aligns with the corresponding element in $\mathbf{u}$. Now, let's establish the following definition:

$$S_{\alpha\lambda}(\mathbf{u}) = \begin{bmatrix} u_0 \\ u_1 - \alpha\lambda\frac{u_1}{||u_1||_2} \\ \cdots \\ u_n - \alpha\lambda\frac{u_n}{||u_n||_2} \end{bmatrix}. \tag{27}$$

An alternative representation of Equation (26) is achieved by transforming it into:

$$S_{\alpha\lambda}(\mathbf{u}) = (1 + \frac{(1-\alpha)\lambda}{||\mathbf{z}||_2}) \begin{bmatrix} z_0 \\ z_1 \\ \cdots \\ z_n \end{bmatrix}. \tag{28}$$

If we apply the L2 norm to both sides as follows:

$$\begin{aligned} ||S_{\alpha\lambda}(\mathbf{u})||_2 &= (1 + \frac{(1-\alpha)\lambda}{||\mathbf{z}||_2}) \cdot ||\mathbf{z}||_2 \\ \Rightarrow ||\mathbf{z}||_2 &= ||S_{\alpha\lambda}(\mathbf{u})||_2 - (1-\alpha)\lambda. \end{aligned} \tag{29}$$

Upon substituting Equation (29) into Equation (28), we obtain:

$$\mathbf{z} = (1 - \frac{(1-\alpha)\lambda}{||S_{\alpha\lambda}(\mathbf{u})||_2}) \cdot S_{\alpha\lambda}(\mathbf{u}) \tag{30}$$

To summarize:

$$\begin{aligned} \mathbf{prox}_h(\mathbf{u}) &= \begin{cases} 0 & \textbf{if } ||S_{\alpha\lambda}||_2 \leq (1-\alpha)\lambda \\ (1 - \frac{(1-\alpha)\lambda}{||S_{\alpha\lambda}||_2}) \cdot S_{\alpha\lambda}(\mathbf{u}) & \textbf{if } ||S_{\alpha\lambda}||_2 > (1-\alpha)\lambda \end{cases} \\ &= (1 - \frac{(1-\alpha)\lambda}{\max\left(||S_{\alpha\lambda}||_2, (1-\alpha)\lambda\right)}) \cdot S_{\alpha\lambda}(\mathbf{u}) \end{aligned} \tag{31}$$

A.2 Discussion and the Proof of Theorem 5.1.

The identifiability condition for the unrolled, temporally extended DAG $\in \mathbb{R}^{d\times T}$, which includes all variables across all time steps, has been established in the work [12]. Specifically, it assumes that the edges within the graph $\mathcal{G}^*$ remain constant over time. Furthermore, it assumes that for any given window $X \in \mathbb{R}^{d\times w}$, where w represents the window's width, the distribution $\mathbf{P}_X$ over the variables within this window stays invariant across different timesteps. This implies that the conditional distribution $\mathbf{P}(x_{i,t}|\mathbf{PA}(x_{i,t}))$ for any variable x_i is independent of the time index t. The subset of all DAGs that can be segmented in this manner into a directed sequence of a repeating subgraph or window is defined as $\mathcal{D}_s$. This segmentation is based on the fact that repetition of the same conditional distributions and edges over time corresponds to stationary or fixed dynamics. The *IGC* methodology, notable for its time-windowed framework, provides the flexibility to detect causal relationship in $\mathbf{W}_{:,j,t',e_0}^{i,t}$ between any pair of variables $(x_{i,t}, x_{j,t'})$ with a given time lag $p = t' - t$ across any DAG or within any time window. When the assumptions outlined in Theorem 5.1. holds, Theorem

1 from [2] becomes applicable, ensuring that our learned graph $\hat{\mathcal{G}}$ is $\mathcal{I}$-Markov equivalent to $\mathcal{G}^*$. Additionally, given that $\hat{\mathcal{G}} \in \mathcal{D}$, by invoking the Theorem 3.2 from [12], $\hat{\mathcal{G}}$ is $(\mathcal{I}, \mathcal{D})$-Markov equivalent to $\mathcal{G}^*$. Since the true Granger causal structure is originally derived from $\mathcal{G}^*$, by establishing $\hat{\mathcal{G}}$ is $(\mathcal{I}, \mathcal{D})$-Markov equivalent to $\mathcal{G}^*$, we have addressed the challenges related to the identifiability in Granger causality and mitigate the concerns associated with accurately determining causal relationships within the framework of Granger causality. Theorem 1 from [2] and Theorem 3.2 from [12] operate under the implicit assumption that, for each intervention k, the ground truth interventional target $\mathbf{I}^*_{e_k}$ is precisely known. This assumption, however, does not always available in real-world scenarios. To address this discrepancy, we propose an extension to Theorem 3.2 from [12] that accommodates unknown interventional targets. In this context, as our proposed *IGC* method, the interventional targets $\mathcal{I}$ are learned in a manner similarly to how the graph $\mathcal{G}$ is determined. We are now ready to prove our Theorem 5.1.

Proof. Leveraging Theorem 2 from [2], we address scenarios where $\mathcal{I} \neq \mathcal{I}^*$. The core concept of the proof is that $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) > \mathcal{S}(\mathcal{G}, \mathcal{I})$ whenever $\mathcal{G} \notin (\mathcal{I}^*, \mathcal{D})$-MEC($\mathcal{G}^*$) or when $\mathcal{I} \neq \mathcal{I}^*$. For the sake of clarity, we define:

$$\eta(\mathcal{G}, \mathcal{I}) := \inf_{\phi} \sum_{k \in [K]} D_{KL}(P_{e_k} || F^{e_k}_{\mathcal{G}\mathcal{I}\phi}). \tag{32}$$

Lemma A.1. *Let $i \in V$ and $A \subset V \backslash \{i\}$, if $(p^1, p^2) \notin \mathcal{Z}(i, A)$ and both p^1 and p^2 are strictly positive, then:*

$$\inf_{(f^1, f^2) \in \mathcal{Z}(i,A))} D_{KL}(p^1 || f^1) + D_{KL}(p^2 || f^2) > 0. \tag{33}$$

Case 1. Let $\mathbb{I}$ represent the set of all intervention sets $\mathcal{I}$ for which there is at least one intervention $k_0 \in [K]$ and one variable $i \in [d]$ such that i is included in the true intervention set $\mathbf{I}^*_{k_0}$ but is not included in $\mathbf{I}_{k_0}$. Assuming $\mathcal{I} \in \mathbb{I}$ and considering $\mathcal{G}$ as an arbitrary DAG, the principle of $\mathcal{I}^*$-faithfulness implies that

$$P_{e_0}(x_i | \mathbf{PA}(x_i)) \neq P_{e_{k_0}}(x_i | \mathbf{PA}(x_i)). \tag{34}$$

It also means $(P_{e_0}, P_{e_{k_0}}) \notin \mathcal{Z}(i, \mathbf{PA}(i))$, where,

$$\mathcal{Z}(i, A) := \{(f^1, f^2) | f^1(x_i | x_A) = f^2(x_i | x_A) \text{ and } f^1, f^2 > 0\}. \tag{35}$$

Given that $i \notin \mathbf{I}_{k_0}$, it follows from the definition provided in Equation (7) that, for all values of ϕ,

$$\begin{aligned} &F^{e_0}_{\mathcal{G}\mathcal{I}\phi}(x_i | \mathbf{PA}(x_i)) = F^{e_{k_0}}_{\mathcal{G}\mathcal{I}\phi}(x_i | \mathbf{PA}(x_i)) \\ &\textbf{i.e.}(F^{e_0}_{\mathcal{G}\mathcal{I}\phi}, F^{e_{k_0}}_{\mathcal{G}\mathcal{I}\phi}) \in \mathcal{Z}(i, \mathbf{PA}(i)). \end{aligned} \tag{36}$$

The following holds and for all ϕ we have $(F^{e_0}, F^{e_{k_0}}) \in \mathcal{Z}(i, \mathbf{PA}(i))$ due to Lemma A.1:

$$\begin{aligned} \eta(\mathcal{G}, \mathcal{I}) &\geq \inf_{\phi} D_{KL}(P_{e_0} || F^{e_0}_{\mathcal{G}\mathcal{I}\phi}) + D_{KL}(P_{e_{k_0}} || F^{e_{k_0}}_{\mathcal{G}\mathcal{I}\phi}) \\ &\geq \inf_{(F^{e_0}, F^{e_{k_0}}) \in \mathcal{Z}(i, \mathbf{PA}(i))} D_{KL}(P_{e_0} || F^{e_0}) + D_{KL}(P_{e_{k_0}} || F^{e_{k_0}}) \\ &> 0. \end{aligned} \tag{37}$$

For $\min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\} \geq 0$, then $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) > \mathcal{S}(\mathcal{G}, \mathcal{I})$. Let us define $\mathbb{S} := \{(\mathcal{G}, \mathcal{I}) \in DAG \times \mathbb{I} | \min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\} < 0\}$. To prove that $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}) > 0$ for all $(\mathcal{G}, \mathcal{I}) \in \mathbb{S}$, we need to choose $\lambda_{\mathcal{G}}, \lambda_{\mathcal{I}} > 0$ small enough. Choosing $\lambda_{\mathcal{G}} + \lambda_{\mathcal{I}} < \min_{(\mathcal{G}, \mathcal{I}) \in \mathbb{S}} \frac{\eta(\mathcal{G}, \mathcal{I})}{-\min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}}$ since:

$$\begin{aligned} \lambda_{\mathcal{G}} + \lambda_{\mathcal{I}} &< \min_{(\mathcal{G}, \mathcal{I}) \in \mathbb{S}} \frac{\eta(\mathcal{G}, \mathcal{I})}{-\min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}} \\ &\Leftrightarrow \lambda_{\mathcal{G}} + \lambda_{\mathcal{I}} < \frac{\eta(\mathcal{G}, \mathcal{I})}{-\min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}}; \forall (\mathcal{G}, \mathcal{I}) \in \mathbb{S} \\ &\Leftrightarrow -(\lambda_{\mathcal{G}} + \lambda_{\mathcal{I}}) \min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\} < \eta(\mathcal{G}, \mathcal{I}) \\ &\Leftrightarrow 0 < \eta(\mathcal{G}, \mathcal{I}) + (\lambda_{\mathcal{G}} + \lambda_{\mathcal{I}}) \min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}, \end{aligned} \tag{38}$$

then we have:

$$\begin{aligned} &0 < \eta(\mathcal{G}, \mathcal{I}) + (\lambda_{\mathcal{G}} + \lambda_{\mathcal{I}}) \min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}; \forall (\mathcal{G}, \mathcal{I}) \in \mathbb{S} \\ &\leq \eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{G}}(|\mathcal{G}| - |\mathcal{G}^*|) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|) \\ &= \mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}). \end{aligned} \tag{39}$$

From this point forward, we can assume that $\mathbf{I}^*_k \subset \mathbf{I}_k$ for all $k \in [K]$, and this assumption is valid because any deviation from this condition would fall under Case 1.

Lemma A.2. *Given the assumptions outlined in Theorem 5.1:*

$$\begin{aligned} \mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) = \inf_{\phi} &\sum_{k \in [K]} D_{KL}(P_{e_k} || F^{e_k}_{\mathcal{G}\mathcal{I}\phi}) \\ &+ \lambda_{\mathcal{G}}(|\mathcal{G}| - |\mathcal{G}^*|) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|). \end{aligned} \tag{40}$$

Case 2. Let $\bar{\mathbb{I}} := \{\mathbf{I} | \mathbf{I}^*_k \subset \mathbf{I}_k \forall k \} \text{and} [\exists (k_0, i) \text{ s.t. } i \in \mathbf{I}_{k_0} \text{ and } i \notin \mathbf{I}^*_{k_0}]\}$. Given that $\mathcal{I} \in \bar{\mathbb{I}}$ and $\mathcal{G}$ is a DAG, it becomes evident that $|\mathcal{I}| > |\mathcal{I}^*|$. If $|\mathcal{G}| \geq |\mathcal{G}^*|$, then $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}) > 0$ by Lemma A.2. Define a set $\bar{\mathbb{S}} := \{(\mathcal{G}, \mathcal{I}) \in DAG \times \bar{\mathbb{I}} | |\mathcal{G}| < |\mathcal{G}^*|\}$. To prove that $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}) > 0$ for all $(\mathcal{G}, \mathcal{I}) \in \bar{\mathbb{S}}$, we need to choose $\lambda_{\mathcal{G}}$ small enough. Choosing $\lambda_{\mathcal{G}} < \min_{(\mathcal{G}, \mathcal{I}) \in \bar{\mathbb{S}}} \frac{\eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|)}{|\mathcal{G}| - |\mathcal{G}^*|}$ since:

$$\begin{aligned} \lambda_{\mathcal{G}} &< \frac{\eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|)}{|\mathcal{G}| - |\mathcal{G}^*|}; \forall (\mathcal{G}, \mathcal{I}) \in \bar{\mathbb{S}} \\ &\Leftrightarrow \lambda_{\mathcal{G}}(|\mathcal{G}| - |\mathcal{G}^*|) < \eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|) \\ &\Leftrightarrow 0 < \eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{G}}(|\mathcal{G}| - |\mathcal{G}^*|) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|), \end{aligned} \tag{41}$$

then we have:

$$\begin{aligned} &0 < \eta(\mathcal{G}, \mathcal{I}) + \lambda_{\mathcal{G}}(|\mathcal{G}| - |\mathcal{G}^*|) + \lambda_{\mathcal{I}}(|\mathcal{I}| - |\mathcal{I}^*|); \forall (\mathcal{G}, \mathcal{I}) \in \bar{\mathbb{S}} \\ &= \mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}). \end{aligned} \tag{42}$$

In the scenarios described by Case 1 and Case 2, all instances where $\mathcal{I} \neq \mathcal{I}^*$ are accounted for. Consequently, this leads to the conclusion that $\mathcal{I}$ must be equal to $\mathcal{I}^*$. By noting that $\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}) = \mathcal{S}_{\mathcal{I}^*}(\mathcal{G}^*) - \mathcal{S}_{\mathcal{I}^*}(\mathcal{G})$, we can employ the same steps as [12] to prove that $\hat{\mathcal{G}} \in (\mathcal{I}^*, \mathcal{D}) - \text{MEC}(\mathcal{G}^*)$. □