Causal Structure Learning for Latent Intervened Non-stationary Data

Chenxi Liu [1] **Kun Kuang** [1]

Abstract

Causal structure learning can reveal the causal mechanism behind natural systems. It is well studied that the multiple domain data consisting of observational and interventional samples benefit causal identifiability. However, for non-stationary time series data, domain indexes are often unavailable, making it difficult to distinguish observational samples from interventional samples. To address these issues, we propose a novel *Latent Intervened Non-stationary learning (LIN)* method to make *the domain indexes recovery process* and *the causal structure learning process* mutually promote each other. We characterize and justify a possible faithfulness condition to guarantee the identifiability of the proposed LIN method. Extensive experiments on both synthetic and real-world datasets demonstrate that our method outperforms the baselines on causal structure learning for latent intervened non-stationary data.

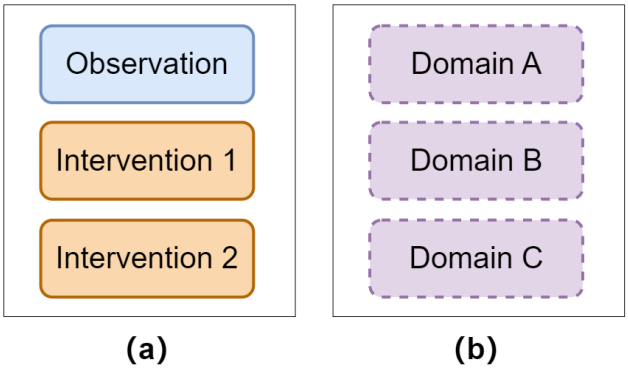

Figure 1. Comparison of input data in different settings. (a) the multi-domain setting: Samples are partitioned according to their provided domain indexes (solid borders); observational samples are available and are distinguishable from interventional samples (different colors). (b) *latent-intervened* setting. Samples are mixed without domain indexes (dashed borders); observational samples are not distinguishable and are technically possible to be absent (same color).

1. Introduction

Causal structure learning is one fundamental problem in causal inference and aims to learn the causal mechanism/structure among variables from observational data. The causal structure is often represented by a Directed Acyclic Graph (DAG). One benefit of the causal structure is to help understand the complex mechanisms in the real world, like diseases (Shen et al., 2020) or earth systems (Runge et al., 2019a); and can also benefit downstream tasks (Pearl, 2000; 2010).

Traditional methods for causal structure learning include constraint-based (Spirtes et al., 1995; 2000; Zhang, 2008) , score-based (Chickering, 2002) , and model-based (Shimizu et al., 2006; Hoyer et al., 2008) approaches. They assume data is independent and identically distributed, which could be violated in heterogeneous/non-stationary cases, resulting in spurious correlation or non-independent noise. Recently, a new line of research on *the multi-domain methods* shows valuable information in such cases could aid causal structure learning (Yang et al., 2018; Ghassami et al., 2018; Brouillard et al., 2020; Huang et al., 2020; Perry et al., 2022). They learned the consistent causal structure from data with multiple different domains, instructed by domain indexes. Formally, a domain index $e_i \in [K]$ indicates the domain generating sample $x_i \in \mathbb{R}^d$.

Unfortunately, such domain indexes are usually unavailable for non-stationary temporal data in reality, which limits their applications. There are attempts to fill this gap. CD-NOD (2020) assumed time indexes can approximate domain indexes with smoothness assumption; Faria et al (2022) assumed the Dirichlet process (DP) over domain indexes to treat ELBO instead. However, most of the existing methods impose additional assumptions on the dynamic over domain indexes, like smoothness or certain types of process; or require samples from observational distribution to be included and can be distinguished from other domains, limiting their applications in more general cases. Therefore, as illustrated

[1]College of Computer Science and Technology, Zhejiang University, Zhejiang, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

in Figure 1, learning causal structures for *latent-intervened* non-stationary data is still an open problem.

In this paper, we focus on causal structure learning task for *latent intervened* non-stationary data, which emphasizes on following aspects, compared with the previous task:

1. **Generality.** Special assumptions can limit the dynamic of domain indexes, e.g. smoothness or certain processes. These assumptions can induce bias once violated.

2. **Samples.** Most methods (Yang et al., 2018; Brouillard et al., 2020) require samples in the domain with no intervention to be distinguishable from samples in intervened domains. However, in reality, these samples can be hard to identify or can be not included.

3. **Explainability.** Information about which and how variables' distributions change can be useful in reality. Huang et. al.(2020) visualized the rapidness of change over time. However, It is still hard to identify a certain variable's distribution shifting at different time points.

For this new task, we propose a novel *Latent Intervened Non-stationary learning (LIN)* method with considering *both learning causal structure and recovering domain indexes*. Specifically, we answer the following questions. The first question is, ***when and how well can we recover domain indexes without assumption on their dynamic?*** We show that is possible (up to Kullback–Leibler divergence and permutations) during the causal structure learning process. This result allows our method to be applicable in more general cases. And the second question is, ***when and how well can we learn causal structure without access to domain indexes?*** This situation is more realistic, however, its faithfulness condition is less studied in the current existing literature. We characterize and justified a possible faithfulness condition which guarantees *$\mathcal{I}$-Markov Equivalence Class ($\mathcal{I}$-MEC)*. Later we shall see these questions benefit each other. For explainability, for different time points, we identify their domain indexes and corresponding intervention targets to show whether one variable' conditional distribution changes and provide information for downstream tasks.

The main contributions of this paper can be summarized as follows:

1. This paper formalizes the casual structure learning task for latent intervened non-stationary data. The domain indexes that are associated with different interventional distributions are unavailable. Samples from observational distribution may be indistinguishable from other domains.

2. This paper proposes a theoretical-guaranteed score-based semi-parametric method, called *Latent Intervened Non-stationary learning (LIN)* to both recover latent domain indexes and learn causal structure.

3. **For domain indexes**, this paper shows that they are possible to be recovered (up to Kullback–Leibler divergence and permutations) without assumptions on their dynamic. **For causal structure**, this paper characterizes a new $\mathcal{I}$-faithfulness condition to guarantee $\mathcal{I}$-Markov Equivalence Class ($\mathcal{I}$-MEC).

4. We provide extensive experiments with existing methods on both synthetic data and real-world data. For real-world data, we compared results in recovering the Pacific Walker circulation from a climate dataset.

2. Related Works

Structure learning from multiple domains A series of parametric works (Ghassami et al., 2018; Yang et al., 2018; Brouillard et al., 2020) treat data from different domains as intervention distribution. i.e. the distribution of an intervention distribution is a result of changing some conditional distribution from observational distribution. For each intervention distribution, one augmented node is added to the original graph with links from this node to those nodes whose conditional distributions are intervened. The identification requires assumptions on distributions.

There are also non-parametric approaches that introduce one single surrogate variable to indicate which variable changes across different domains. Mooij et al. (2020) introduce a framework to learn pooled data from different contexts. Some works (Huang et al., 2020; Perry et al., 2022) utilize distribution shifting by kernel embedding or counting pairs of domains where conditional distribution changes. These methods allow non-parametric conditional independence tests between distributions.

Structure learning for non-stationary data One research line is to adapt the FCI algorithm to the structural vector-autoregressive (SVAR) process (Malinsky & Spirtes, 2018; 2019; Gerhardus & Runge, 2020). These methods can learn both instantaneous and time-lag relations with latent confounders while assuming the non-stationarity is due to the partial observation (i.e. the violation of the causal sufficiency assumption) on a stationary process with no distribution shifting.

Another research line is to utilize distribution shifting. There exists an attempt to model the coefficients of the linear SVAR model it-selves as another set of linear SVAR models (Huang et al., 2019). It imposes a specific structure over non-stationarity and is restricted to the linear case. Recently some methods (Zhang et al., 2017; Huang et al., 2020) use

time indexes as a surrogate variable to detect modules changing with time. They showed that distribution shifting can benefits structure learning. They assume that latent variables are deterministic smooth functions of time, and the method doesn't explicitly split data into different clusters.

3. Non-stationarity from Latent Interventions

In this section, we describe the data generating process and discuss how non-stationarity emerges from latent interventions.

Data Generating Process Consider a time series $\{x_t\}_{t=1}^T$, where $x_t \in \mathbb{R}^d$. t is a time index. The time series is assumed to follow a causal graph $\mathcal{G}(V, E)$. V is the set of variables, and E is the set of edges. If we only consider instantaneous relation, then $|V| = d$. Here we also consider time-lag relation up to $t - P + 1$ so that $V = \{x_t^{(1)}, \ldots, x_t^{(d)}, \ldots, x_{t-P+1}^{(1)}, x_{t-P+1}^{(d)}\}$, and $|V| = dP$. For $i \in [d]$, $\mathrm{Pa}_i \subset V$ is the set of direct causes for $x_t^{(i)}$ in $\mathcal{G}$.

If no intervention is applied, the joint distribution for x_t has the following factorization:

$$p^{(\emptyset)}(x_t) = \prod_{j \in [d]} p^{(\emptyset)}_{j|\,\mathrm{Pa}_j}. \tag{1}$$

- $p^{(\emptyset)}(x_t)$ called observational distribution, where $\emptyset$ indicates no variable is intervened.
- Samples generated according to $p^{(\emptyset)}$ are called observational samples.
- $p^{(\emptyset)}_{j|\,\mathrm{Pa}_j}$ refers to the conditional distribution of $x_t^{(j)}$ given Pa_i.

Domains As in the multi-domain literature (Mooij et al., 2020; Huang et al., 2020; Perry et al., 2022), samples are divided into different domains by their generating processes. Specifically, each sample x_t is assigned a domain index $e_t = k \in [K]$. Its corresponding distribution is

$$p^{(I_k)}(x_t) = \prod_{j \in I_k} p^{(I_k)}_{j|\,\mathrm{Pa}_j} \prod_{j \notin I_k} p^{(\emptyset)}_{j|\,\mathrm{Pa}_j}. \tag{2}$$

- I_k is the intervention targets in the k-th domain, and $p^{(I_k)}_{j|\,\mathrm{Pa}_j} \neq p^{(\emptyset)}_{j|\,\mathrm{Pa}_j}$.
- If $I_k \neq \emptyset$, $p^{(I_k)}(x_t)$ is called interventional distribution; samples are called interventional samples.
- We sometime use $p^{(k)}$ to refer $p^{(I_k)}$ for short.

In this paper, domain indexes are assumed to be latent. To make sure the above setting is well-defined, a common assumption in the multi-domain literature is employed: Requiring the set of intervention targets $\mathcal{I} := \{I_k \mid k \in [K]\}$ to be conservative, as stated in Definition 3.1.

Definition 3.1. An intervention targets set $\mathcal{I}$ over d variables $\{1, 2, \ldots, d\}$ is said to be *conservative* if for any $j \in [d]$, there exist $I \in \mathcal{I}$ such that $j \notin I$.

Intuitively, when $\mathcal{I}$ is conservative, each conditional distribution of $p^{(\emptyset)}$ is presented as a part of the interventional distributions in some domains. Therefore, the dataset contains enough information about the observational distribution. For more details, please see Appendix B.

Non-stationarity Here, samples are associated with two types of indexes: time indexes and domain indexes. Time indexes allow sample dependency through time-lag relations in the causal graph. Domain indexes allow samples to follow different conditional distributions sharing the common causal graph. In some realistic situations, domain indexes can be hard to acquire. Therefore, the underlying distribution shifting cannot be captured, yielding a non-stationarity in the aspect of the data generating process.

4. Latent Intervened Non-stationary Learning

In this section, we propose a method to both learn the underlying causal graph and domain indexes. Previous literature finds that domain indexes are helpful to identify the causal graph. We highlight that its converse is also true: the causal graph can reduce the searching space of possible data distribution, which is also helpful in recovering domain indexes.

4.1. Model and Algorithm

Sometimes we will emphasize the ground truth by adding a star symbol, for example, $\mathcal{G}^*$ for the causal graph, $\mathcal{I}^*$ for the set of intervention targets, and $\mathcal{E}^*$ for the sample partition based on domain indexes. Note that the $\mathcal{E}^*$ is unique up to permutation over its clusters.

Model To learn domain indexes, a hyper-parameter N_c is required to specify the number of domains. Our method will divide samples into at most N_c clusters, where each cluster is to estimate one domain. The distribution in the c-th cluster is estimated by a corresponding parametric model $f^{(c)}$

$$\begin{aligned} f^{(c)}(x_t) &:= \prod_{j=1}^{d} f^{(c)}_{j|\,\mathrm{Pa}_j}, \\ c &\in \{1, 2, \ldots, N_c\}. \end{aligned} \tag{3}$$

Note that in Equation (3), Pa_j is now an *estimation* for causal parents Pa_j^* of node j in true causal graph $\mathcal{G}^*$.

An intuitive approach to further expand Equation (3) is to parameterize the intervention targets $\hat{I}_c$ and observational

distribution $f^{(\emptyset)}$ directly. However, we argue this is not a good way. In our problem setting, domain indexes are not available. As a consequence, we have no valid initial samples to feed to $f^{(\emptyset)}$. Another reason is that when $\mathcal{I}$ is conservative, information about observational distribution has already been included in domains' distributions, modeling an additional $f^{(\emptyset)}$ would introduce redundant parameters, and has a negative influence on the learning process.

Therefore, we expand Equation (3) in an indirect way. Particularly, we introduce a lower triangular matrix $M_j \in \{0,1\}^{N_c \times N_c}$ for each node $j \in [d]$. $M_{jkl} = 1$ means the k-th and the l-the cluster shared a common conditional distribution, i.e. $f^{(k)}_{j|\,\mathrm{Pa}_j} = f^{(l)}_{j|\,\mathrm{Pa}_j}$. Each row of M_j contains exactly one element with value 1 and others with value 0. To be specific:

$$\begin{aligned} f^{(c)}_{j|\,\mathrm{Pa}_j} &:= \prod_{\ell=1}^{k} \left(\tilde{f}^{(c)}_{j|\,\mathrm{Pa}_j}\right)^{M_{jc\ell}}, \\ \tilde{f}^{(c)}_{j|\,\mathrm{Pa}_j} &:= \tilde{f}^{(c)}\big(x_t^{(j)} | \theta = \mathrm{NN}_{k,j}(\mathrm{Pa}_j; \phi_k)\big), \\ & c \in \{1, 2, \ldots, N_c\}, \end{aligned} \tag{4}$$

where $\tilde{f}^{(c)}$ is a pre-defined distribution family, whose parameters are determined by a neural network NN based on variables' parents' values. In the technical aspect, each M_j should have a unique representation, to do this, an additional regularization term is employed:

$$g(M) = \sum_{j=1}^{d} \sum_{c=1}^{N_c} \sum_{\ell=1}^{c} M_{jc\ell} \cdot \ell \tag{5}$$

By using $g(M)$, different slots on the same line in each M_j have different priorities, which makes M_j to be the representation unique that minimizes $g(M)$ (rows are fixed, no permutation on rows).

Score We learn the proposed model by maximizing the score we present here.

$$\begin{aligned} \mathcal{S}(\mathcal{G}, M, \mathcal{E}) := \sup_{\phi} \frac{1}{T} \sum_{c=1}^{N_c} \sum_{t \in \mathcal{E}_e} \log f^{(c)}(x_t) \\ - \lambda |\mathcal{G}| - \lambda_M g(M) \end{aligned} \tag{6}$$

The first term is the averaged log-likelihood over data;The second and the third terms are penalty terms with positive small coefficients $\lambda, \lambda_M > 0$. ϕ stands for parameters in neural networks. $|\mathcal{G}|$ is number of edges in graph.

Here we highlight that M can be seen as an estimation of the set of intervention targets $\mathcal{I}$. We cannot write out $\mathcal{I}$ only because we don't know *which* one is from observational distribution. By definition, M estimates the similarity among domains for each conditional distribution. Therefore, given $\mathcal{I}$ is conservative, M also contains the similarity between each conditional distribution with its counterpart in observational distribution. So we sometimes refer M by notation $\mathcal{I}$ to emphasize that we will use the similarity information related to observational distribution.

It is also reasonable to use the notation $|\mathcal{I}|$ to refer to $g(M)$. When a node $j \in [d]$ is added to an intervention targets I_c of a domain, the $g(M)$ will increase. Before adding the node, the row $M_{jc,\cdot}$ will select one column, say the ℓ-th, i.e. $M_{jc\ell} = 1$. By conservative assumption, there are multiple non-zero items in the ℓ-th column. Under the faithfulness condition 5.5 which we shall discuss later, the ℓ-th column will be split into two columns, leading to one additional non-zero column in M_j. Therefore, weights of some items originally in the ℓ-th column will increase in $g(M)$.

It is worth noting that the score we propose in Equation (6) is *not* a trivial analogy to the case where domain indexes are given (Brouillard et al., 2020), where samples from the same domain were firstly aggregated to compute domain-specific log-likelihoods. Directly applying their score in this setting introduces a biased term since we don't know the correct partition. The induced biased term is difficult to tackle and would influence the identifiability.

Algorithm 1 Latent Intervention Learning

Input: non-stationary data X, hyper-parameter N_c, ρ_0
Randomly Initialization:$\{\mathcal{G}\}, M = \{M_j\}_{j=1}^{d}, \{\mathcal{E}\}$, $\mathcal{E}$ is a sample partition.
Initialization Lagrange multiplier: $\alpha \leftarrow 0, \rho \leftarrow \rho_0$
repeat
  Solve sub-problem $L^{\rho}(\mathcal{G}, \mathcal{I}, \alpha \mid \mathcal{E})$ until it converges
  Update α and ρ based on (Zheng et al., 2018)
  Update $\{\mathcal{E}\}$ as described.
until hold-out loss converged and $h = 0$.
Output: estimated graph $\hat{\mathcal{G}}$, the set of matrices $\{\hat{M}_j\}_{j=1}^{d}$, and clusters $\hat{\mathcal{E}}$

Algorithm We present our algorithm in Algorithm 1. The algorithm received the non-stationary dataset X without domain indexes, and a hyper-parameter N_c is given to suggest how many clusters the algorithm should consider. Then the samples would be initially separated in N_c clusters randomly. The causal graph is initialized to be a fully connected graph, and each M_j is randomly initialized. Two Lagrange multipliers are employed, and are initialized as $\alpha := 0$, and $\rho := \rho_0$. And we take ρ_0 as 10^{-8} in practice.

Then the algorithm would basically go through an optimization process in order to maximize the score under the constraint that the instantaneous relation in causal graph $\mathcal{G}$ forms a directed acyclic graph. Zheng et al. (Zheng et al., 2018) characterize DAG constraint by calculating trace for

the matrix exponential to adjacency matrix. For example, let $\mathcal{G}_0$ be the adjacency graph for instantaneous relation. The constrain condition requires $\operatorname{Tr} e^{\mathcal{G}_0} = 0$.

An augmented Lagrange optimization process would be used. We solve a series of sub-problem

$$\begin{aligned} &\min_{\mathcal{G},\mathcal{I},\alpha} L^\rho(\mathcal{G},\mathcal{I},\alpha|\mathcal{E}) \\ &\quad L^\rho(\mathcal{G},\mathcal{I},\alpha \mid \mathcal{E}) \\ &:= -\mathcal{S}(\mathcal{G},\mathcal{I},\mathcal{E}) + \alpha|h(\mathcal{G})| + \frac{\rho}{2}|h(\mathcal{G})|^2 \end{aligned} \tag{7}$$

where $h(\mathcal{G})$ is Non-DAG penalty for instantaneous links (Zheng et al., 2018). Define as $h(\mathcal{G}) := \operatorname{Tr} e^{\mathcal{G}_0}$, where $\mathcal{G}_0$ is the adjacency graph for instantaneous relation. ρ and α would be initialized by ρ_0 and α_0 and would be updated according to the augmented Lagrange optimization process. Data partition $\mathcal{E}$ would be updated by clustering process between sub-problems.

4.2. Recovering domain indexes

As shown in Algorithm 1, the estimation for domain indexes $\mathcal{E}$ would be updated after each sub-problem. Each sample's domain index is estimated by

$$\arg\max_{k\in[N_c]} \log f^{(k)}(x_t). \tag{8}$$

Following previous works, we employ the Gumbel-softmax trick (Jang et al., 2017) to keep back-propagation in practice.

In the process of causal structure identification, the domain indexes would be gradually recovered. The following statement is a corollary of Theorem 5.10, which is presented here to informally explain how far we can go on Recovering domain indexes.

Corollary 4.1. *Given the condition stated in Theorem 5.10, if the score proposed in Equation (6) is optimized, then samples in each cluster would approach one of the true domains asymptotically in the sense of Kullback–Leibler divergence.*

5. Analysis on Identifiability

In this section, we discuss the Identifiability of our proposed method and the assumptions it requires. Some assumptions are employed in the case where domain indexes are given (Brouillard et al., 2020): Assumption 5.7, Assumption 5.8.

5.1. Preliminary

Relations among interventional distributions In this paper, we assume domain indexes are unavailable, therefore, it is hard to test whether observational samples are included in a dataset, not to mention distinguishing them from others. As suggested by related work, it can be helpful to *pretend* one domain is observational.

Definition 5.1. (Yang et al., 2018). J-observation targets set. For a set of interventional distribution $\{p^{(I)}\}_{I\in\mathcal{I}}$ with a intervention target set $\mathcal{I}$, where $\emptyset \in \mathcal{I}$ may not hold. For one specific target $J \in \mathcal{I}$, we relabeled the original target set to $\tilde{\mathcal{I}}_J$ with corresponding interventional distribution set $\{\tilde{p}^{(I)}\}_{I\in\tilde{\mathcal{I}}_J}$ by following rules:

- for $I \in \mathcal{I}$, if $I = J$, then relabel I as $\emptyset$ in $\tilde{\mathcal{I}}_J$.
- for $I \in \mathcal{I}$, if $I \neq J$, then relabel I as $\tilde{I}_J := I \cup J$
- $\tilde{p}_J^{(\emptyset)} := p^{(J)}$, and $\tilde{p}_J^{(\tilde{I}_J)} := p^{(I)}$

Then we have:

$$\begin{aligned} p^{(I)}(x_t) &= p^{(\tilde{I}_J)}(x_t) \\ &= \prod_{j\in\tilde{I}_J} p^{(I)}_{j|\,\mathrm{Pa}_j} \prod_{j\notin\tilde{I}_J}^{d} p^{(J)}_{j|\,\mathrm{Pa}_j} \\ &= \prod_{j\in\tilde{I}_J} [\tilde{p}_J^{(\tilde{I}_J)}]_{j|\,\mathrm{Pa}_j} \prod_{j\notin\tilde{I}_J}^{d} [\tilde{p}_J^{(\emptyset)}]_{j|\,\mathrm{Pa}_j} \end{aligned} \tag{9}$$

Causal graph with intervention Previous works use the notion of $\mathcal{I}$-DAG to graphically represent intervention information by introducing additional nodes and edges.

Definition 5.2. (Yang et al., 2018) Given a DAG $\mathcal{G}(V, E)$ which is a causal graph, and an intervention target set $\mathcal{I}$, whose each element $I \in \mathcal{I}$ is a subset of V, indicating which variables are intervened by that intervention. $\mathcal{I}$-DAG $\mathcal{G}^{\mathcal{I}}$ is the augmented causal graph with node set $V \cup \{\zeta_I | I \in \mathcal{I}\}$ and edge set $E \cup \{\zeta_I \to j | I \in \mathcal{I}, j \in V, j \in I\}$

Note that the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ for a J-observation targets set $\tilde{\mathcal{I}}_J$ is defined in same way.

$\mathcal{I}$-Markov Equivalence class Given a graph $\mathcal{G}$ and a set of intervention targets $\mathcal{I}$, the set of all multi-domain distributions which it can express, is denoted by $\mathcal{M}_{\mathcal{I}}(\mathcal{G})$. We say $\mathcal{G}_1$ and $\mathcal{G}_2$ are *$\mathcal{I}$-Markov Equivalence Class ($\mathcal{I}$-MEC)* if and only if $\mathcal{M}_{\mathcal{I}}(\mathcal{G}_1) = \mathcal{M}_{\mathcal{I}}(\mathcal{G}_2)$. (Yang et al., 2018)

Definition 5.3. $\mathcal{M}_{\mathcal{I}}(\mathcal{G}) := \{\{p^{(I)}\}_{I\in\mathcal{I}} | \forall I, J \in \mathcal{I} : p^{(I)} \in \mathcal{M}(\mathcal{G})$ and $p^{(I)}_{j|\mathrm{Pa}_j} = p^{(J)}_{j|\mathrm{Pa}_j}, \forall j \notin I \cup J\}$

where $\mathcal{M}(\mathcal{G})$ is the collection of strictly positive densities which are Markov to $\mathcal{G}$.

Important Notations We use $A \perp\!\!\!\perp_{\mathcal{G}} B|C$ to represent that node A and node B are d-separated given node C in the causal graph $\mathcal{G}$;and use $A \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} B|C$ to represent that node A and node B are d-separated given node C in the

augmented causal graph $\mathcal{G}^{\mathcal{I}}$. For random variables, we use $X_A \perp\!\!\!\perp_{p^{(I)}} X_B | X_C$ to represent that variable X_A and X_B are independent conditioned on variable X_C under distribution $p^{(I)}$.

5.2. Faithfulness condition with latent domains

Is it necessary to propose a new faithfulness condition, given the one that has been proposed in the multi-domain literature (Yang et al., 2018; Brouillard et al., 2020)? For latent intervened non-stationary data, it also lost the information about *which domain represents the observational one for each node's conditional distribution*. In this case, it can be ambiguous to determine interventional targets. As a consequence, augmented nodes in estimated $\mathcal{I}$-DAG graph can point to wrong variables, then it would contain wrong v-structures, and therefore some edges with wrong direction in the original non-augmented causal graph.

Based on Yang et al.'s result (2018), Brouillard et al.(2020) formalized $\mathcal{I}$-*faithfulness* condition. They showed that this condition is required to learn $\mathcal{I}$-MEC by maximizing the score they designed.

Definition 5.4. $\mathcal{I}$-faithfulness assumption (Brouillard et al., 2020)

1. For disjoint subsets $A, B, C \subset V$, if X_A and X_B are d-connected in graph $\mathcal{G}$ conditioning on X_C, then $X_A \not\perp\!\!\!\perp_{p^{(\emptyset)}} X_B | X_C$
2. For any disjoint $A, C \subset V$ and $I \in \mathcal{I}$, $p^{(I)}_{A|C} = p^{(\emptyset)}_{A|C}$ implies $A \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}$

To show the above condition 5.4's problem in latent intervened data, consider two different domains $I, J \in \mathcal{I}$ with a node $j \in I \cap J$ whose conditionals satisfy

$$p^{(I)}_{j|\,\mathrm{Pa}_j} = p^{(J)}_{j|\,\mathrm{Pa}_j} \neq p^{(\emptyset)}_{j|\,\mathrm{Pa}_j}. \tag{10}$$

In this case, for latent intervened data, there are ambiguous explanations about whether node j should be an intervention target, which motivates our following new faithfulness condition 5.5.

Definition 5.5. Faithfulness for latent domains.

1. For disjoint subsets $A, B, C \subset V$, if X_A and X_B are d-connected in graph $\mathcal{G}$ conditioning on X_C, then exist $I \in \mathcal{I}$, such that $X_A \not\perp\!\!\!\perp_{p^{(I)}} X_B | X_C$
2. For any disjoint $A, C \subset V$ and $I \neq J$ in $\mathcal{I}$,

$$[\tilde{p}_J^{(\tilde{I}_J)}]_{A|C} = [\tilde{p}_J^{(\emptyset)}]_{A|C}$$

implies

$$A \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{I}_J} | C \cup \zeta_{-\tilde{I}_J}$$

The second condition in Definition 5.5 rejects the case stated in Equation (10), and requires the interventions to be diverse enough to avoid repeating conditionals.

In the last part of this subsection, we characterize the relation of these two sets of faithfulness conditions by the following proposition, where (C2) is from Definition 5.4, (C4) is from Definition 5.5; (C1) is necessary for (C4); and (C3) is necessary for (C2).

Proposition 5.6. *Suppose $\{p^{(I)}\}_{I\in\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, and $\mathcal{I}$ is conservative, then: conditions (C1) and (C2) hold if and only if conditions (C3) and (C4) hold.*

- *(C1) For any disjoint $A, C \subset V$ and $I \neq J$ in $\mathcal{I}$, if $[\tilde{p}_J^{(\tilde{I}_J)}]_{A|C} = [\tilde{p}_J^{(\emptyset)}]_{A|C}$, then there exists $\tilde{L}_J \in \tilde{\mathcal{I}}_J$ and $\tilde{L}_J \neq \emptyset$, such that $A \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{L}_J} | C \cup \zeta_{-\tilde{L}_J}$*
- *(C2) For any disjoint $A, C \subset V$ and $I \in \mathcal{I}$, $p^{(I)}_{A|C} = p^{(\emptyset)}_{A|C}$ implies $A \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}$*
- *(C3) For any disjoint $A, C \subset V$ and $I \in \mathcal{I}$, if $p^{(I)}_{A|C} = p^{(\emptyset)}_{A|C}$, then there exists $J \in \mathcal{I}$ and $J \neq \emptyset$ such that $A \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_J | C \cup \zeta_{-J}$*
- *(C4) For any disjoint $A, C \subset V$ and $I \neq J$ in $\mathcal{I}$, $[\tilde{p}_J^{(\tilde{I}_J)}]_{A|C} = [\tilde{p}_J^{(\emptyset)}]_{A|C}$ implies $A \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{I}_J} | C \cup \zeta_{-\tilde{I}_J}$*

5.3. Main result for identifiability

Let $\mathcal{G}^*$ be the ground truth causal graph, M^* characterizing the ground truth intervention target set $\mathcal{I}^*$, and $\mathcal{E}^*$ is the correct partition based on domain index. Pa_j^* is the set of parents for node j in ground truth graph $\mathcal{G}^*$.

Assumption 5.7. For the ground truth distribution of data $\{p^{(I)}\}_{I\in\mathcal{I}^*} \in M_{\mathcal{I}^*}(\mathcal{G}^*)$, and they have finite entropy. For the ground truth intervention target set $\mathcal{I}^*$ assume it is conservative, i.e. for any $j \in [p]$, there is an $I \in \mathcal{I}$, such that $p^{(I)}_{j|\mathrm{Pa}_j^*} = p^{(\emptyset)}_{j|\mathrm{Pa}_j^*}$

Assumption 5.8. For model density, assume it is strictly positive, and the true density can be expressed by Equation (2).

Assumption 5.9. The ground truth distribution $\{p^{(I)}\}_{I\in\mathcal{I}^*}$ is faithful to $\mathcal{G}^*$ and $\mathcal{I}^*$ according to Definition 5.5

Theorem 5.10. *With Assumption 5.7, 5.8, and 5.9, in addition to $N_c \geq K$, each cluster has enough data, and the penalty coefficients in Equation (6) is sufficiently small, it holds asymptotically that for any estimation $(\hat{\mathcal{G}}, \hat{\mathcal{I}}, \hat{\mathcal{E}})$*

$$\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*, \mathcal{E}^*) > \mathcal{S}(\hat{\mathcal{G}}, \hat{\mathcal{I}}, \hat{\mathcal{E}})$$

, if $\hat{\mathcal{G}}_{\hat{\mathcal{I}}}$ is not an $\mathcal{I}$-Markov Equivalence to ${\mathcal{G}^*}_{\mathcal{I}^*}$, or any cluster in $\hat{\mathcal{E}}$ is close to none of the domains in the sense of Kullback–Leibler divergence.

Table 1. Results in synthetic data driven by Brownian motion

ERDOS-RENYI BROWNIAN DIM METHOD	1D ACCU	1D RECALL	1D F1	1D SHD	1D SID	2D ACCU	2D RECALL	2D F1	2D SHD	2D SID
PCMCI+	0.7800	0.5385	0.5600	11	14	0.8400	0.6923	0.6923	8	5
VAR-LINGAM	0.6800	**1.0000**	0.6190	16	**0**	0.6400	**1.0000**	0.5909	18	**0**
DYNOTEARS	0.7800	0.6923	0.6207	11	21	0.8200	0.6154	0.6400	9	23
SVAR-FCI	0.7600	0.4615	0.5000	12	16	0.8400	0.6154	0.6667	8	6
LATENT PCMCI	0.8000	0.6154	0.6154	10	12	0.7800	0.6923	0.6207	11	4
CPF_SAEM	0.5400	0.5385	0.3784	23	16	0.6800	0.8462	0.5789	16	9
TIME-LAG CD-NOD	0.6600	0.4615	0.4138	17	21	0.6400	0.5385	0.4375	18	22
DCDI	0.7200	0.9231	0.6316	14	2	0.7400	**1.0000**	0.6667	13	**0**
LIN (OURS)	**0.9200**	**1.0000**	**0.9286**	**2**	**0**	**0.8800**	**1.0000**	**0.8125**	**6**	**0**
BARABASI–ALBERT BROWNIAN DIM METHOD	1D ACCU	1D RECALL	1D F1	1D SHD	1D SID	2D ACCU	2D RECALL	2D F1	2D SHD	2D SID
PCMCI+	0.7000	0.6000	0.5455	15	23	0.7800	0.5333	0.5926	11	22
VAR-LINGAM	0.5400	0.8667	0.5306	23	6	0.4800	0.8000	0.4800	26	16
DYNOTEARS	0.7000	0.6667	0.5714	15	21	0.6800	0.7333	0.5789	16	30
SVAR-FCI	0.7200	0.3333	0.4167	14	30	0.7600	0.4667	0.5385	12	17
LATENT PCMCI	0.7800	0.7333	0.6667	11	11	0.8200	0.8000	0.7273	9	11
CPF_SAEM	0.6400	0.4667	0.4375	18	31	0.6200	0.5333	0.4571	19	20
TIME-LAG CD-NOD	0.6600	0.4000	0.4138	17	35	0.6600	0.4667	0.4516	17	34
DCDI	0.7200	0.6000	0.5625	14	25	0.8600	0.9333	0.8000	7	7
LIN (OURS)	**0.9800**	**1.0000**	**0.9677**	**1**	**0**	**0.8800**	**1.0000**	**0.8333**	**6**	**0**

In Theorem 5.10 above, none of *Enough data* nor *Sufficient small* implies additional assumption. *Enough data* means we are using sample average to estimate the expectation of the proposed score, which is common in related literature. Hence the estimated score $\mathcal{S}(\mathcal{G}, \mathcal{I}, \mathcal{E})$ could be seen as a random variable. If N_c is too large, the algorithm may have no enough data, which leads to a large variance and consequently has a negative influence on identifiability. *Sufficient small* means the coefficients λ and λ_M in Equation (6) should be sufficient small in the sense of Equation (34) and Equation (35), which also occurs in one of our most related work (Brouillard et al., 2020). The value of λ and λ_M can be selected by empirical criterion stated in the next section.

6. Experiments

Empirical Criterion for Hyper-parameter Selection We introduce a rule of thumb to select hyper-parameter N_c. The first term is the ratio of the parameters' number to sample size; the second term is the loss (i.e. negative score) over hold-out data. When N_c is too large, the first term would be dominant, and such N_c would not be selected. When the sample size is sufficient, the second term would be dominant and one could select N_c mainly based on likelihood. Other criteria could be used based on background knowledge, for example, use averaged negative log-likelihood.

$$\text{criterion} = \frac{\#\text{param}}{\#\text{sample}} + \text{eval.loss} \tag{11}$$

Metrics We use these metrics across all experiments: accuracy, recall, F1 score , SHD, and SID . *Accuracy* is the rate that one model correctly predicts the existence or absence of edges in the ground truth graph. *Recall* is the proportion of edges in the ground truth graph that are correctly detected by a model. *F1 score* is the harmonic mean between precision and recall, where precision is the ratio of correctly detected edges to all predicted edges. *SHD* means Structural Hamming Distance, which is the number of edges' states that are mistake predicted. *SID* means Structural Intervention Distance, which measures the distance between the estimated graph and true graph in terms of their corresponding causal inference statements (Peters & Bühlmann, 2015).

Baseline Methods We compare the proposed method with three groups of methods. The first group is **for the stationary data** : (1) PCMCI+ (Runge, 2020): use CI tests with optimized conditioning sets. (2) VAR-LiNGAM (Hyvärinen et al., 2010): use ICA (Hyvärinen & Oja, 2000) to non-Gaussian data. (3) dyNoTears (Pamfil et al., 2020): a score-based method with continuous optimization. The second group is **for latent confounders** : (1) SVAR-FCI (Malinsky & Spirtes, 2018): adapt FCI (Spirtes et al., 1995) to time-series. (2) Latent PCMCI (Gerhardus & Runge, 2020): adapt PCMCI+ to time-series. The third group is **for distribution shifting** : (1) CPF-SAEM: linear structure with state-space model (Huang et al., 2019). (2) CD-NOD: two-stage non-parametric method (Huang et al., 2020). (3) DCDI:

score-based parametric method (Brouillard et al., 2020).

6.1. Synthetic Data

The data are generated based on the combination of two types of graphs, and two types of stochastic processes that are employed to drive the latent interventions, and for each process type, we test for two different instances.

Graph type we consider Erdős–Rényi graph (ER graph) and Barabási–Albert (BA graph). For the ER graph, each edge is generated with a fixed probability. For the BA graph nodes are added one by one and edges from each new node to existing nodes are generated with a probability that is proportional to the existing edge one node has.

$$p_i = \frac{k_i + 1}{\sum_j (k_j + 1)} \tag{12}$$

Stochastic process At each time point, one domain index would be sampled according to multinomial distribution. we compared cases when each domain indexes are equally weighted and when they are not. In unequal case, the sampling probability is proportional to $\begin{bmatrix}1 & 2 & \dots & K\end{bmatrix}$.

Another random process we use is Brownian motion. we condition i-dimension Brownian motion in $\mathbb{R}^i$, where $i \in \{1, 2\}$. We restrict the motion inside the box $[-M, M]^i$ by simply applying $x \mapsto (x \mod 2M) - M$, and each quadrant represents one domain.

Result In Table 1, the system is under a Brownian motion, moving in a 1-dimensional or 2-dimensional rectangle centered at the original point. Each quadrant represents one domain. In Table 3 (see appendix), the domain of each time point is decided by multinomial sampling, which is evenly or unevenly distributed over possible domains.

We observe that (1) For CD-NOD, its smoothness assumption over domain indexes is violated in these cases. As a result, its performance is lower than the group for latent confounders; and has a gap with DCDI receiving random domain indexes. CPF-SAEM has similar result. These observations support our argument that *assumptions on the dynamic of domain indexes limit the generality*. (2) In the group for latent confounders, Latent PCMCI has a relatively better result. However, it doesn't consistently outperform methods for stationary data. Therefore, it is not sufficient to handle non-stationarity by simply regarding domain indexes as a latent confounder because it *ignores the information carried by distribution shifting*, as pointed out by the multi-domain methods in recent years; and is supported by the result of our LIN method. (3) Our LIN method outperforms methods in the group for distribution shifting and the other two groups. This empirical fact corroborates our theoretical analysis of our method's identifiability over latent domain indexes associated with different interventions in different situations.

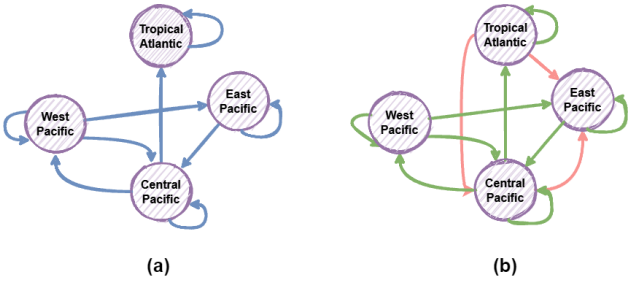

Figure 2. Result illustration on Pacific Walker Circulation Data. (a) ground truth (b) LIN Method. Green: Correctly detected. Red: detected but not exist.

6.2. Real World Data

Pacific Walker Circulation (PWC) dataset Pacific Walker Circulation (PWC) describes wind movement among the Pacific. It is an important and well-understood topic in climate science (Runge et al., 2019b). The dataset is provided by Copernicus Climate Change Service information (Hersbach et al., 2023). The ground truth causal graph and related coordinates of latitude and longitude are adopted from Eldhose et al. (2022). The dataset contains daily records of surface pressure or temperature in four regions for 20 years since 2001. The ground truth graph contains loops, so we treat it as a summary graph and therefore SID metric is not applicable here.

Illustration We draw the result in Figure 2. The left part is the ground truth graph for Pacific Walker Circulation, which is explained by Runge et. al.(2019b); The right part is from our LIN method. The method correctly recovered a structure that is a non-trivial super-set of the ground truth.

Result Results are summarized in Table 2. We could observe that (1) Among the three groups of baseline methods, the group for distribution shifting has better results in accuracy and F_1 score. (2) In this time-series data, smoothness is a relatively reasonable assumption, so we could observe CD-NOD and CPF-SAEM have fair good results. CPF-SAEM produced the second best results in all metrics. (3) The Pacific system is evolving under intervention by surrounding unknown factors. Hence it is reasonable to expect our LIN method outperforms others. (4) VAR-LiNGAM also produced the highest recall, however, it predicted all time-lag edges, leading to the highest recall but a trivial summary graph with very limited information and the highest SHD. (5) Combined with synthetic data, we see that

Table 2. Results in Pacific Walker Circulation data. *(SID is not applicable for summary graph with loops)*

METHOD	ACCU	RECALL	F1	SHD
PCMCI+	0.6250	0.5556	0.6250	6
VAR-LINGAM	0.5625	**1.0000**	0.7200	7
DYNOTEARS	0.6875	0.4444	0.6154	5
SVAR-FCI	0.6250	0.5556	0.6250	6
LATENT PCMCI	0.6875	0.6667	0.7059	5
CPF_SAEM	0.7500	0.7778	0.7778	4
TIME-LAG CD-NOD	0.6875	0.6667	0.7059	5
DCDI	0.6875	0.6667	0.7059	5
LIN (OURS)	**0.8125**	**1.0000**	**0.8571**	**3**

our LIN method outperforms baseline methods in different dynamics of domain indexes (i.e. smooth or stochastic), which supports its improvement in generality.

6.3. Clustering Performance

We also evaluated the model's ability for clustering in synthetic data. Markham et al.(2022) discussed several kernel-based methods for clustering data with respect to their inherent structure which are included as baseline methods.

Result Results are shown in Table 4 in Appendix. The first row is driven by multinomial sampling, and the second row is driven by Brownian motion. Our method has a higher ARI score in each case. These results validate our LIN's ability to recover hidden domain indexes during the causal structure learning procedure. It is needed to note that in non-stationary data, the baseline methods' i.i.d. assumption is violated, which is another possible reason for their lower score.

Another perspective is to compare ARI scores for each method across different cases. We could observe that the same method can have different ARI scores in different graph types or stochastic processes. One possible reason is that the Kullback–Leibler divergence between two domains varies with cases. This is because we generate conditional distribution randomly in the data generating process. If two domain distributions have smaller Kullback–Leibler divergence, then they are more difficult to be distinguished.

6.4. Hyper-parameter Analysis

We provide results with different combinations of hyper-parameters in Table 6 and Table 7. Due to page limitation, detailed analysis is presented in Appendix A.4.

7. Conclusion

In this paper, we investigated the causal structure learning task for *latent intervened* non-stationary time series data , which is a more realistic and more difficult multi-domain setting *with the absence of domain indexes*. The previous faithfulness condition has a problem in this situation because it can have ambiguous interpretations for intervention targets, leading to incorrect v-structures and edges with false directions.

For causal identifiability, we characterize a new faithfulness condition to guarantee the recovery of domain indexes and the identification of causal structure to $\mathcal{I}$-MEC. For the learning algorithm, we propose a novel *Latent Intervened Non-stationary Learning (LIN)* method to identify causal structure up to $\mathcal{I}$-MEC and to recover domain indexes up to Kullback–Leibler divergence. In agreement with theoretical analysis, our LIN method outperforms previous methods in synthetic and real-world datasets while without assumptions over domain indexes.

Possible future works include estimating the exact number of hyper-parameter N_c, analyzing sample efficiency of the method, and exploring non-parametric methods for this task.

Potential Negative Societal Impact

Our methods require faithfulness conditions to learn causal structure, and there is no result guarantee about selection bias in this paper. Hence potential users should be familiar with the background of their data, and be cautious about making unfair decisions due to biased datasets.

Source Code

Our code is available at LIN2023 on GitHub. For codes for baseline methods, please refer to causal-learn, lingam, tigramite, dynotears, and CPF_SAEM.

Acknowledgement

This work was supported in part by National Key Research and Development Program of China (2021YFC3340300), National Natural Science Foundation of China (62006207, 62037001, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Project by Shanghai AI Laboratory (P22KS00111) and Zhejiang Province Natural Science Foundation (LQ21F020020). The real-world climate data are provided by the Copernicus program. We are also grateful for the helpful comments provided by anonymous reviewers.

References

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable Causal Discovery from Interventional Data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877. Curran Associates, Inc., 2020.

Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002.

Eldhose, E., Chauhan, T., Chandel, V., Ghosh, S., and Ganguly, A. R. Robust Causality and False Attribution in Data-Driven Earth Science Discoveries. *Arxiv*, 2022.

Faria, G. R. A., Martins, A., and Figueiredo, M. A. T. Differentiable causal discovery under latent interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 253–274. PMLR, 11–13 Apr 2022.

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

Ghassami, A., Kiyavash, N., Huang, B., and Zhang, K. Multi-domain causal structure learning in linear systems. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6ad4174eba19ecb5fed17411a34ff5e6-Paper.pdf.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. Era5 hourly data on single levels from 1940 to present, 2023. DOI: 10.24381/cds.adbb2d47 (Accessed on January 2023).

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Huang, B., Zhang, K., Gong, M., and Glymour, C. Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2901–2910. PMLR, 2019.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes. *Journal of Machine Learning Research*, 2020.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4): 411–430, 2000. ISSN 0893-6080.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Malinsky, D. and Spirtes, P. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pp. 23–47. PMLR, August 2018.

Malinsky, D. and Spirtes, P. Learning the Structure of a Nonstationary Vector Autoregression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 2986–2994. PMLR, April 2019.

Markham, A., Das, R., and Grosse-Wentrup, M. A Distance Covariance-based Kernel for Nonlinear Causal Clustering in Heterogeneous Populations. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 542–558. PMLR, June 2022. ISSN: 2640-3498.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 2020.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. Dynotears: Structure learning from time-series data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605. PMLR, 26–28 Aug 2020.

Pearl, J. *Causality: Models, reasoning, and inference*. Cambridge University Press, New York, NY, US, 2000.

Pearl, J. Causal Inference. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, pp. 39–58. PMLR, 2010.

Perry, R., von Kügelgen, J., and Schölkopf, B. Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis. In *Advances in Neural Information Processing Systems*, 2022.

Peters, J. and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3): 771–799, March 2015.

Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397. PMLR, 03–06 Aug 2020.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10:2553, 2019a.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, November 2019b. Publisher: American Association for the Advancement of Science.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Shen, X., Ma, S., Vemuri, P., and Simon, G. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. *Scientific Reports*, 10:2975, 2020.

Shimizu, S., Hoyer, P. O., Hyvarinen, A., and Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, pp. 28, 2006.

Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pp. 499–506, 1995.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Wang, Y., Solus, L., Yang, K., and Uhler, C. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yang, K., Katcoff, A., and Uhler, C. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1347–1353, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

A. Additional Experiment

A.1. Experiment Result on synthetic data

Table 3. Results in synthetic data driven by multinomial sampling

Erdos-Renyi Multinomial Weight	Even					Uneven				
method	accu	recall	F1	SHD	SID	accu	recall	F1	SHD	SID
PCMCI+	0.8800	0.6923	0.7500	6	10	0.7600	0.5385	0.5385	12	14
VAR-LiNGAM	0.5400	0.8462	0.4889	23	5	0.6000	0.9231	0.5455	20	1
dyNoTears	0.8600	0.7692	0.7407	7	21	0.8600	0.6923	0.7200	7	10
SVAR-FCI	0.8200	0.5385	0.6087	9	15	0.8400	0.6154	0.6667	8	12
latent PCMCI	0.8600	0.6923	0.7200	7	8	0.8800	0.8462	0.7857	6	2
CPF_SAEM	0.6400	0.6154	0.4706	18	13	0.5800	0.6154	0.4324	21	15
time-lag CD-NOD	0.7800	0.5385	0.5600	11	21	0.7200	0.6154	0.5333	14	21
DCDI	0.8200	0.9231	0.7273	9	3	0.7800	0.9231	0.6857	11	1
LIN (ours)	**0.9600**	**1.0000**	**0.9286**	**2**	**0**	**0.9200**	**1.0000**	**0.8667**	**4**	**0**
Barabasi–Albert Multinomial Weight	Even					Uneven				
method	accu	recall	F1	SHD	SID	accu	recall	F1	SHD	SID
PCMCI+	0.6800	0.4667	0.4667	16	25	0.5600	0.2667	0.2667	22	27
VAR-LiNGAM	0.5000	0.7333	0.4681	25	14	0.5200	0.8000	0.5000	24	12
dyNoTears	0.4000	**1.0000**	0.5000	30	20	0.6600	0.8000	0.5854	17	28
SVAR-FCI	0.7600	0.4000	0.5000	12	24	0.7400	0.2667	0.3810	13	27
latent PCMCI	0.7200	0.4667	0.5000	14	23	0.7600	0.6000	0.6000	12	17
CPF_SAEM	0.7000	0.5333	0.5161	15	28	0.6400	0.4000	0.4000	18	33
time-lag CD-NOD	0.6800	0.3333	0.3846	16	37	0.6600	0.2667	0.3200	17	36
DCDI	0.7600	0.8000	0.6667	12	11	0.6600	0.8000	0.5854	17	11
LIN (ours)	**0.9400**	**1.0000**	**0.9091**	**3**	**0**	**0.9000**	**1.0000**	**0.8571**	**5**	**0**

A.2. Experiment Result on clustering

Table 4. Clustering Results in synthetic data.

	ER even		ER uneven		BA even		BA uneven	
Method	RI	ARI	RI	ARI	RI	ARI	RI	ARI
K-means	0.6296	0.1919	0.7750	0.5147	0.5520	0.0140	0.5738	0.0858
Dep-Con	0.6053	0.1244	0.5411	0.0146	0.6776	0.2746	0.6634	0.2773
Poly-kernel	0.6053	0.1244	0.5366	0.0035	0.6603	0.2435	0.6747	0.3034
RBF	0.6255	0.1849	0.5922	0.1228	0.5527	0.0221	0.5630	0.0637
LIN (ours)	0.9519	**0.8919**	0.9057	**0.7988**	0.9167	**0.8126**	0.7303	**0.4184**
	ER 1d		ER 2d		BA 1d		BA 2d	
Method	RI	ARI	RI	ARI	RI	ARI	RI	ARI
K-means	0.8581	0.7161	0.6662	0.2107	0.5042	0.0084	0.5782	0.0136
Dep-Con	0.5019	0.0038	0.5478	0.0052	0.5020	0.0040	0.6301	0.1449
Poly-kernel	0.5073	0.0147	0.5851	0.0088	0.5015	0.0030	0.6338	0.1548
RBF	0.6704	0.3409	0.6641	0.2071	0.5051	0.0101	0.5777	0.0157
LIN (ours)	0.9996	**0.9992**	0.7919	**0.4481**	0.8123	**0.6247**	0.7463	**0.3000**

A.3. Experiment Result on Highly Heterogeneous Real-world Data

Dataset We analysis experiment results on Sachs (2005) dataset. This dataset consists of 11 measurements about protein signals over 7466 cells under different experimental conditions. In this example, different domains correspond to treatments on cells by different combinations of chemical reagents. To build a highly heterogeneous data, we simply ignore the domain indexes information and treat them as mixed data.

Methods This data is not a time series. We adapting our LIN method by setting $P = 1$, i.e. considering no time-lag relations. And we replace baseline methods used in time-series situations with their non-time-lag counterparts.

Discussion The experiment result are shown in Table 5. Note that some assumptions of our method are violated in this dataset because chemical reagents can affect receptor enzymes which are not included in measured molecules (Wang et al., 2017), which means the dataset is not causal sufficient and interventions can be applied to latent variables.

Although the assumptions are not fully satisfied, our LIN method still yields at least the second-best results in all 5 metrics among baselines. Table 5 also suggests that LiNGAM may be also a fair good method to treat mixed heterogeneous data, as it attained the highest recall and lowest SID. DCDI method yields third-best results in all 5 metrics, which indicates that our domain recovering process is still more helpful than random partition for the causal structure learning task.

Table 5. Results in mixed Sachs data without domain indexes

METHOD	ACCU	RECALL	F1	SHD	SID
PC	0.7603	0.2105	0.2162	29	90
LINGAM	0.6364	**0.6316**	0.3529	44	**47**
LINEAR-NOTEARS	0.8099	0.1579	0.2069	23	92
FCI	0.7355	0.2632	0.2381	32	89
LATENT PCMCI	0.6694	0.1053	0.0909	40	101
CPF_SAEM	0.7438	0.3684	0.3111	31	96
CD-NOD	0.7603	0.2105	0.2162	29	90
DCDI	0.7603	0.3684	0.3256	29	85
LIN (OURS)	**0.8182**	0.5263	**0.4762**	**22**	57

A.4. Discussion on hyper-parameters

In this section, we discuss the effect of hyper-parameters with results shown in Table 6 and Table 7, and also show the efficiency of our proposed empirical criterion.

In Table 6, we perform our method with different combinations of hyper-parameters on synthetic data where domain indexes are generated by multinomial sampling over 3 domains. The first column indicates whether the sampling weights of domains are evenly distributed. In Table 7, domain indexes are generated by Brownian motions. The first column indicates the number of dimensions of this motion. For 1-d Brownian motion, there are 2 domains, For 2-d Brownian motion, there are 4 domains.

The first row in each table indicates the type of the true causal graph, where ER graph means Erdős–Rényi graph and BA graph means Barabási–Albert graph. Under each graph type, 6 matrices are presented. The first three metrics are usual accuracy, recall, and F1 score for directed edge detection, and the following two are Structural Hamming Distance (SHD) as well as Structural Intervention Distance (SID). The last one is our hyper-parameter selection criterion, described in section 6. This criterion is agnostic to any evaluation metric, and the lowest criterion is preferred. As one could observe from the tables, this criterion can select a combination yielding reasonable results. Users of our method can adjust the criterion based on their background knowledge of datasets.

In Table 6 and Table 7, we analysis the combinations of two important hyper-parameters:

- PNT: the coefficient of penalty terms,
- N_c: the number of clusters assigned to the model.

Discussion Fixing PNT $= 1E-8$, we discuss the influence of N_c. (1) One can verify that whenever N_c is larger than the true value, our method can yield highly competitive (best results except for very few cases) result in all 5 metrics compared with the other 8 baseline methods. Therefore, our method is not very sensitive to N_c. (2) Although the identifiability still holds when N_c is larger than the true value, an extremely large N_c can raise issues in optimization due to limited sample size. In addition, a too-small N_c may fail to capture the non-stationarity inside data. In these cases, it is possible to yield non-optimal results as shown in Table 6 and Table 7. (3) When the true number of domains is unknown, it is necessary to adjust N_c and PNT by utilizing the empirical criterion. One can see that the empirical criterion selects reasonable results in all cases.

Now we discuss the influence of PNT by comparing PNT $= 1E-8$ and PNT $= 1E-4$ (and use N_c selected by empirical criterion). In general, a smaller PNT is preferred with respect to identifiability, as one can observe that PNT $= 1E-8$ is better in most cases. On the other hand, a larger PNT can encourage the optimization process to converge faster (but not necessarily better) by reducing searching space. In addition, using a larger PNT can be interpreted as encouraging sparsity.

Table 6. hyper-parameters analysis on synthetic data driven by multinomial sampling

			ER graph						BA graph					
weights	PNT	N_c	accu	recall	F1	SHD	SID	sel	accu	recall	F1	SHD	SID	sel
Even	1E-8	2	0.9200	1.0000	0.8667	4	0	-3.2353	0.7800	0.8667	0.7027	11	7	-2.2741
		3	0.9600	1.0000	0.9286	2	0	**-3.6975**	0.9400	1.0000	0.9091	3	0	**-3.0597**
		4	0.9400	1.0000	0.8966	3	0	-3.6333	0.8800	1.0000	0.8333	6	0	-2.3980
		5	0.9800	1.0000	0.9630	1	0	-3.6642	0.8400	0.8667	0.7647	8	11	-2.6892
	1E-4	2	0.9200	1.0000	0.8667	4	0	-3.3149	0.8800	0.9333	0.8235	6	1	-2.4696
		3	0.9400	1.0000	0.8966	3	0	**-3.6148**	0.8800	0.8667	0.8125	6	9	**-3.1210**
		4	0.9200	1.0000	0.8667	4	0	-3.5125	0.8400	0.9333	0.7778	8	5	-2.8086
		5	0.9200	1.0000	0.8667	4	0	-3.1754	0.8600	0.8667	0.7879	7	11	-2.2944
Unev	1E-8	2	0.8600	0.9231	0.7742	7	1	-2.5793	0.8000	0.8667	0.7222	10	9	-2.8719
		3	0.9200	1.0000	0.8667	4	0	**-3.3498**	0.9200	0.9333	0.8750	4	4	-3.3998
		4	0.9000	0.9231	0.8276	5	1	-3.1910	0.9000	0.8667	0.8387	5	8	-3.2579
		5	0.8400	0.9231	0.7500	8	1	-3.1059	0.9000	1.0000	0.8571	5	0	**-3.4644**
	1E-4	2	0.8400	0.9231	0.7500	8	1	-2.6234	0.8600	0.9333	0.8000	7	3	-2.7123
		3	0.9200	0.9231	0.8571	4	1	**-3.3880**	0.8400	0.9333	0.7778	8	6	-3.2681
		4	0.9200	1.0000	0.8667	4	0	-3.1917	0.8800	0.9333	0.8235	6	7	**-3.4001**
		5	0.9200	0.9231	0.8571	4	1	-3.2738	0.8600	0.9333	0.8000	7	4	-3.1593

Table 7. hyper-parameters analysis on synthetic data driven by Brownian motion

			ER graph						BA graph					
dimension	PNT	N_c	accu	recall	F1	SHD	SID	sel	accu	recall	F1	SHD	SID	sel
1d	1E-8	2	0.9600	1.0000	0.9286	2	0	**-3.2053**	0.9000	0.9333	0.8485	5	5	-4.1417
		3	0.8400	1.0000	0.7647	8	0	-3.0937	0.9800	1.0000	0.9677	1	0	**-4.2656**
		4	0.8600	1.0000	0.7879	7	0	-3.1295	0.8600	0.8667	0.7879	7	8	-3.5480
		5	0.9200	1.0000	0.8667	4	0	-3.0565	0.8800	0.8667	0.8125	6	7	-3.7747
	1E-4	2	0.9200	1.0000	0.8667	4	0	-3.1107	0.9800	1.0000	0.9677	1	0	-4.0558
		3	0.8000	1.0000	0.7222	10	0	-2.9197	0.8800	0.9333	0.8235	6	3	-3.7044
		4	0.8400	1.0000	0.7647	8	0	-3.0473	0.9400	0.9333	0.9032	3	5	**-4.1255**
		5	0.8800	1.0000	0.8125	6	0	**-3.3169**	0.9600	0.9333	0.9333	2	5	-3.7997
2d	1E-8	2	0.8000	1.0000	0.7222	10	0	-2.7519	0.7800	0.9333	0.7179	11	10	-1.5107
		3	0.8800	1.0000	0.8125	6	0	-3.2873	0.8800	1.0000	0.8333	6	0	-2.1300
		4	0.8800	1.0000	0.8125	6	0	**-3.2969**	0.8800	1.0000	0.8333	6	0	**-2.1867**
		5	0.9000	1.0000	0.8387	5	0	-3.1764	0.8800	0.9333	0.8235	6	7	-1.4915
	1E-4	2	0.7600	0.9231	0.6667	12	6	-2.3545	0.8000	0.9333	0.7368	10	12	-1.3081
		3	0.9000	1.0000	0.8387	5	0	-3.0087	0.9600	1.0000	0.9375	2	0	**-2.2369**
		4	0.8200	0.9231	0.7273	9	2	**-3.3188**	0.9600	1.0000	0.9375	2	0	-1.9692
		5	0.9000	1.0000	0.8387	5	0	-3.1711	0.9400	0.9333	0.9032	3	3	-1.6238

B. Preliminary Background Knowledge

In this section, we go through some basic concepts in related literature. Proposition B.2 is an extended version of Yang's result (2018) to technically cover the case where $\emptyset \notin \mathcal{I}$, therefore, the proof of Proposition B.2 is not claimed to be a contribution of this paper.

Notation recall $\mathcal{I}$ is a set of intervention targets. $\{p^{(I)}\}_{I\in\mathcal{I}}$ is the corresponding interventional distribution. Meanings of $\mathcal{M}(\mathcal{G})$ and $\mathcal{M}_{\mathcal{I}}(\mathcal{G})$ are in Definition 5.3.

Definition B.1. $\mathcal{I}$-Markov Property

- $p^{(I)} \in \mathcal{M}(\mathcal{G})$ for any $I \in \mathcal{I}$
- For any disjoint subset $A, V \subset V$ and $I \in \mathcal{I}$, if $A \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}$, then $p^{(I)}_{A|C} = p^{(\emptyset)}_{A|C}$

Proposition B.2. *Suppose $\{p^{(I)}\}_{I\in\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, and $\mathcal{I}$ is conservative, then*

- *there uniquely exists $p^{(\emptyset)} \in \mathcal{M}(\mathcal{G})$ such that each $p^{(I)}$ can be factorized like Equation (2)*
- *$\{p^{(I)}\}_{I\in\mathcal{I}}$ satisfies $\mathcal{I}$-Markov properties..*

Proof. If $\emptyset \in \mathcal{I}$ then we are done according to **Yang et. al. 's work(2018)**. Here we consider $\emptyset \notin \mathcal{I}$. The general idea is similar.

Since $p^{(\emptyset)} \in \mathcal{M}(\mathcal{G})$, it is of the form $p^{(\emptyset)} = \prod_{i\in V} p^{(\emptyset)}_{i|\text{Pa}_i}$. For each $i \in V$, since $\mathcal{I}$ is conservative, there exist $I \in \mathcal{I}$ such that $i \notin I$, let $p^{(\emptyset)}_{i|\text{Pa}_i}$ be $p^{(I)}_{i|\text{Pa}_i}$. Such choice is well-defined (and is unique), if $I \neq J \in \mathcal{I}$, and $i \notin I \cup J$, by $\{p^{(I)}\}_{I\in\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, we have $p^{(I)}_{i|\text{Pa}_i} = p^{(J)}_{i|\text{Pa}_i}$.

The first one in $\mathcal{I}$-Markov properties holds by definition of $\mathcal{M}_{\mathcal{I}}(\mathcal{G})$. For the second one, consider any $C \subset V$, and any $j \in V \setminus C$, if there is an $I \in \mathcal{I}$, such that

$$j \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}, \tag{13}$$

then it holds that

$$p^{(I)}_{j|C} = p^{(\emptyset)}_{j|C} \tag{14}$$

As $p^{(\emptyset)}$ exists, such that for each $J \in \mathcal{I}$, if $i \notin J$, then $p^{(J)}_{i|\text{Pa}_i} = p^{(\emptyset)}_{i|\text{Pa}_i}$, where Pa_i means the parents of node i in graph $\mathcal{G}$. We keep this density $p^{(\emptyset)}$ fixed. We define V_{an} as the ancestral set of $\{j\} \cup C$ (including $\{j\} \cup C$); let $B' \subset V_{\text{an}}$ be the nodes in V_{an} that are d-connected to ζ_{-I} conditioning on $C \cup \zeta_{-I}$ in augmented graph $\mathcal{G}^{\mathcal{I}}$; And $A' := V_{\text{an}} \setminus (B' \cup C)$; $j \in A'$.

- For $i \in A'$, by definition, $i \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}$, we have $i \notin I$, which means $p^{(I)}_{i|\text{Pa}_i} = p^{(\emptyset)}_{i|\text{Pa}_i}$. We claim that $\text{Pa}_i \subset A' \cup C$: suppose $\ell \in \text{Pa}_i \setminus C$, if $\ell \in B'$, then there exists a path $i \longleftarrow \ell \longleftarrow \cdots \longleftarrow \zeta_I$ (or $i \longleftarrow \ell \longrightarrow \cdots \longleftarrow \zeta_I$), leading to $i \in B'$ given $\ell \notin C$, contradiction.
- For $i \in B'$, we claim that $\text{Pa}_i \subset B' \cup C$: if not, suppose $\ell \in (\text{Pa}_i \cap A') \setminus C$, if there is a path $i \longrightarrow \cdots \longleftarrow \zeta_I$ conditioning on $C \cup \zeta_I$ in $\mathcal{G}^{\mathcal{I}}$, then through $\ell \longrightarrow i$, we have $\ell \in B'$, contradiction; if the path is $i \longleftarrow \cdots \longleftarrow \zeta_I$, from Equation (13) we know there is no directed path $i \longrightarrow \cdots \longrightarrow j$ without involving nodes in C, and so a common descendant between ℓ and ζ_I is in C, so $\ell \in B'$, contradiction.
- For $i \in C$ such that $\text{Pa}_i \cap A' \neq \emptyset$. Clearly $i \notin I$, otherwise some nodes in A' would also be in B'. So we have $p^{(I)}_{i|\text{Pa}_i} = p^{(\emptyset)}_{i|\text{Pa}_i}$. Similarly, node i has no parents in B'. Hence, $\text{Pa}_i \subset A' \cup C$.
- For $i \in C$ such that $\text{Pa}_i \cap A' = \emptyset$, by definition $\text{Pa}_i \subset V_{\text{An}} \setminus A' = B' \cup C$.

Therefore, for $\hat{I} \in \{\emptyset, I\}$, since $\{p^{(I)}\}_{I\in\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$ and **Lemma A.1 in Yang et. al. 's work(2018)**, we can write:

$$p^{(\hat{I})}(X) = g_1(X_{A'}, X_C) g_2(X_{B'}, X_C; \hat{I}) \prod_{i\in V\setminus V_{\text{an}}} p^{(\hat{I})}_{i|\text{Pa}_i} \tag{15}$$

where

$$g_1(X_{A'}, X_C) = \prod_{i \in A'} p^{(\emptyset)}_{i|\mathrm{Pa}_i} \prod_{i \in C, \mathrm{Pa}_i \cap A' \neq \emptyset} p^{(\emptyset)}_{i|\mathrm{Pa}_i} \tag{16}$$

$$g_2(X_{B'}, X_C; \hat{I}) = \prod_{i \in B'} p^{(\hat{I})}_{i|\mathrm{Pa}_i} \prod_{i \in C, \mathrm{Pa}_i \cap A' = \emptyset} p^{(\hat{I})}_{i|\mathrm{Pa}_i} \tag{17}$$

so

$$p^{(\hat{I})}(X_j, X_C) = \hat{g}_1(X_j, X_C)\hat{g}_2(X_C; \hat{I}) \tag{18}$$

where $\hat{g}_1(X_j, X_C) = \int_{X_{A' \setminus \{j\}}} g_1(X_{A'}, X_C)$, $\hat{g}_2(X_C; \hat{I}) = \int_{X_{B'}} g_2(X_{B'}, X_C; \hat{I})$.

Now we observe that

$$\begin{aligned} & p^{(\hat{I})}_{j|C} \\ =& p^{(\hat{I})}(X_j|X_C) \\ =& \frac{p^{(\hat{I})}(X_j, X_C)}{p^{(\hat{I})}(X_C)} \\ =& \frac{\hat{g}_1(X_j, X_C)\hat{g}_2(X_C; \hat{I})}{\hat{g}_2(X_C; \hat{I}) \int_{X_j} \hat{g}_1(X_j, X_C)} \\ =& \frac{\hat{g}_1(X_j, X_C)}{\int_{X_j} \hat{g}_1(X_j, X_C)} \end{aligned} \tag{19}$$

is invariant with $\hat{I}$, so $p^{(I)}_{j|C} = p^{(\emptyset)}_{j|C}$ □

C. Detailed Proofs for Our Results

Proof for Proposition 5.6. We first consider the *only if* direction. Given conditions (C1) and (C2), (C3) is straightforward. Now we consider (C4). We start with

$$[\tilde{p}^{(\tilde{I}_J)}_J]_{j|C} = [\tilde{p}^{(\emptyset)}_J]_{j|C} \tag{20}$$

for some $C \in V$, $j \in V \setminus C$ and $I \neq J$ in $\mathcal{I}$.

In order to show (C4) holds, we what to show

$$j \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{I}_J} | C \cup \zeta_{-\tilde{I}_J} \tag{21}$$

which means there is no path between j and $\zeta_{\tilde{I}_J}$ in augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ conditioning on $C \cup \zeta_{-\tilde{I}_J}$.

By (C1), there exists $\tilde{L}_J \in \tilde{\mathcal{I}}_J$ and $\tilde{L}_J \neq \emptyset$, such that

$$j \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{L}_J} | C \cup \zeta_{-\tilde{L}_J} \tag{22}$$

If $L = I$ then Equation (21) holds and we are done. If not, since the d-separated condition, for any $\ell \in \tilde{L}_J \coloneqq L \cup J$:

- if $\ell \notin C$, then by Equation (22) path like $j \cdots \longleftarrow \ell \longleftarrow \zeta_{\tilde{L}_J}$ is blocked conditioning on $C \cup \zeta_{-\tilde{L}_J}$ in the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ (path like $j \cdots \longrightarrow \ell \longleftarrow \zeta_{\tilde{L}_J}$ is also blocked conditioning on $C \cup \zeta_{-\tilde{L}_J}$ in the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ since ℓ is not conditioned). In this case, there is no path like $j \cdots \longleftarrow \ell$ conditioning on C in the original graph $\mathcal{G}$: otherwise, such path would also exist in the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ conditioning on $C \cup \zeta_{-\tilde{L}_J}$ (because this can be seen as adding a node $\zeta_{\tilde{L}_J}$ and an edge $\zeta_{\tilde{L}_J} \longrightarrow \ell$ to the original graph $\mathcal{G}$), leading to the violation of Equation (22). Hence, there is also no path like $j \cdots \longleftarrow \ell \longleftarrow \zeta_L$ (or $j \cdots \longleftarrow \ell \longleftarrow \zeta_J$) conditioning on $C \cup \zeta_{-L}$ (or $C \cup \zeta_{-J}$) in the augmented graph $\mathcal{G}^{\mathcal{I}}$.
- if $\ell \in C$, then by Equation (22) path like $j \cdots \longrightarrow \ell \longleftarrow \zeta_{\tilde{L}_J}$ is blocked conditioning on $C \cup \zeta_{-\tilde{L}_J}$ in the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ (path like $j \cdots \longleftarrow \ell \longleftarrow \zeta_{\tilde{L}_J}$ is also blocked conditioning on $C \cup \zeta_{-\tilde{L}_J}$ in the augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ since

ℓ is conditioned). In this case, there is no path like $j \cdots \longrightarrow \ell$ conditioning on C in the original graph $\mathcal{G}$; and there is also no path like $j \cdots \longrightarrow \ell \longleftarrow \zeta_L$ (or $j \cdots \longrightarrow \ell \longleftarrow \zeta_J$) conditioning on $C \cup \zeta_{-L}$ (or $C \cup \zeta_{-J}$) in the augmented graph $\mathcal{G}^{\mathcal{I}}$.

It follows that

$$p^{(\emptyset)}_{j|C} = p^{(L)}_{j|C} = p^{(J)}_{j|C} = p^{(I)}_{j|C} \tag{23}$$

the first equality is by Proposition B.2; the second equality is by Equation (22) with corresponding $\tilde{\mathcal{I}}_J$-Markov properties since $\emptyset \in \tilde{\mathcal{I}}_J$ and $\{p^{(I)}\}_{I \in \tilde{\mathcal{I}}_J} \in \mathcal{M}_{\tilde{\mathcal{I}}_J}(\mathcal{G})$, and the third equality is by Equation (20).

By contrary, if Equation (21) doesn't hold, i.e. such path exists, then by definition, it must exist $\ell \in \tilde{I}_J \coloneqq I \cup J \subset V$ such that j is d-connected to $\tilde{I}_J$ through $\longrightarrow \ell \longleftarrow \zeta_{\tilde{I}_J}$ or $\longleftarrow \ell \longleftarrow \zeta_{\tilde{I}_J}$ with $\ell \in C$ or $\ell \notin C$ respectively. This path should not involve nodes in $\zeta_{-\tilde{I}_J}$ because they are conditioned. If conditioning on $\zeta_{-\tilde{I}_J}$ can open any path, it must serve as a collider, but all nodes in $\zeta_{-\tilde{I}_J}$ have no parent. Therefore, that path between (and including) j and ℓ is made of nodes and edges in $\mathcal{G}$, conditioning on $C \subset V$.

Thus, through ℓ, j is d-connected to ζ_I or ζ_J conditioning on set C in graph $\mathcal{G}^{\mathcal{I}}$, conditioning on $C \cup \zeta_{-I}$ or $C \cup \zeta_{-J}$ respectively. By (C2), $p^{(I)}_{j|C} \neq p^{(\emptyset)}_{j|C}$ or $p^{(J)}_{j|C} \neq p^{(\emptyset)}_{j|C}$, which is contradiction with Equation (23).

Next, we consider the *if* direction. Given conditions (C3) and (C4), (C1) is straightforward. Now we consider (C2). Suppose for any disjoint $A, C \subset V$ and any $j \in A$, $I \in \mathcal{I}$, $p^{(I)}_{j|C} = p^{(\emptyset)}_{j|C}$ we want to show

$$j \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}.$$

By (C3), there exists $J \in \mathcal{I}$ and $J \neq \emptyset$ such that

$$j \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_J | C \cup \zeta_{-J}$$

If $J = I$ we are done. If not, by $\mathcal{I}$-Markov condition, we have $p^{(J)}_{j|C} = p^{(\emptyset)}_{j|C} = p^{(I)}_{j|C}$ i.e. $[\tilde{p}_J]^{(\tilde{I}_J)}_{j|C} = [\tilde{p}_J]^{(\emptyset)}_{j|C}$. By (C4),

$$j \perp\!\!\!\perp_{\mathcal{G}^{\tilde{\mathcal{I}}_J}} \zeta_{\tilde{I}_J} | C \cup \zeta_{-\tilde{I}_J},$$

similar to the previous argument,

$$j \perp\!\!\!\perp_{\mathcal{G}^{\mathcal{I}}} \zeta_I | C \cup \zeta_{-I}$$

□

Proof for Theorem 5.10. We consider ground truth by adding a star notion $\mathcal{G}^*$, $\mathcal{I}^*$, and $\mathcal{E}^*$, and denote $\pi^*_\ell \coloneqq \frac{|\mathcal{E}^*_\ell|}{T}$, $\pi_c \coloneqq \frac{|\mathcal{E}_c|}{T}$,

$\pi_{c\ell} \coloneqq \frac{|\mathcal{E}_c \mathcal{E}_\ell^*|}{T}$. For a possible choice (ignoring penalty terms),

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G}, \mathcal{I}, \mathcal{E}) &= -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \pi_{c\ell} \frac{1}{|\mathcal{E}_e \mathcal{E}_\ell^*|} \sum_{x \in \mathcal{E}_e \mathcal{E}_\ell^*} \left[\log f^{(c)}(x) \right] \\
&\to -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\log f^{(c)} \right] \\
&= -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\sum_{j=1}^{d} \log f^{(c)}_{j|\,\mathrm{Pa}_j} \right] \\
&= -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[-\log \frac{p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}}{f^{(c)}_{j|\,\mathrm{Pa}_j}} + \log p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} \right] \\
&= -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[-\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) - \mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \\
&= -\sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[-\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) \right] - \sup_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[-\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \\
&= \inf_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) \right] + \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \\
&= \inf_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) \right] + \sum_{\ell=1}^{K} \sum_{j=1}^{d} \Big(\sum_{c=1}^{N_c} \pi_{c\ell} \Big) \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \\
&= \inf_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) \right] + \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi^*_{\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right]
\end{aligned}
\tag{24}
$$

Note that $\mathcal{G}, \mathcal{I}, \phi$ are used by model $f^{(k)}$ in Equation (3).

For the score of ground truth (ignoring penalty terms):

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*, \mathcal{E}^*) &\to \inf_{\phi} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi^*_{\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(\ell)}_{j|\,\mathrm{Pa}_j}) \right] + \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi^*_{\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \text{(by Equation (24))} \\
&\to 0 + \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi^*_{\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{H}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*}) \right] \text{(by Assumption 5.8)}
\end{aligned}
\tag{25}
$$

Combining Equation (24) and Equation (25) , we have (considering penalty terms):

$$
\begin{aligned}
&\mathcal{S}(\mathcal{G}^*, \mathcal{I}^*, \mathcal{E}^*) - \mathcal{S}(\mathcal{G}, \mathcal{I}, \mathcal{E}) \\
&= \inf_{\phi} \sum_{c=1}^{N_c} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{c\ell} \mathop{\mathbb{E}}_{x \sim p^{(\ell)}} \left[\mathrm{D_{KL}}(p^{(\ell)}_{j|\,\mathrm{Pa}_j^*} || f^{(c)}_{j|\,\mathrm{Pa}_j}) \right] \\
&\quad + \lambda \big(|\mathcal{G}| - |\mathcal{G}^*| \big) + \lambda_M \big(|\mathcal{I}| - |\mathcal{I}^*| \big)
\end{aligned}
\tag{26}
$$

where $|\mathcal{I}| = g(M)$ and $|\mathcal{I}^*| = g(M^*)$. The first term in Equation (26) is the score term, others are penalty terms.

Suppose there exist a cluster e that is not pure,i.e., exist $a, b \in [K]$ with $a \neq b$ but $\pi_{ea} > 0$ and $\pi_{eb} > 0$. Since they are

different distributions, there exists $j \in [p]$ such that $p^{(b)}_{j|\operatorname{Pa}^*_j} \neq p^{(a)}_{j|\operatorname{Pa}^*_j}$. Then the score term in Equation (26) has lower bound:

$$\begin{aligned}
&\inf_{\phi} \sum_{\ell=1}^{K} \sum_{j=1}^{d} \pi_{e\ell} \mathop{\mathbb{E}}_{x\sim p^{(\ell)}} \mathrm{D}_{\mathrm{KL}}(p^{(\ell)}_{j|\operatorname{Pa}^*_j} || f^{(e)}_{j|\operatorname{Pa}_j}) \\
\geq &\inf_{\phi} \left\{ \pi_{ea} \mathop{\mathbb{E}}_{x\sim p^{(a)}} \left[\mathrm{D}_{\mathrm{KL}}(p^{(a)}_{j|\operatorname{Pa}^*_j} || f^{(e)}_{j|\operatorname{Pa}_j}) \right] + \pi_{eb} \mathop{\mathbb{E}}_{x\sim p^{(b)}} \left[\mathrm{D}_{\mathrm{KL}}(p^{(b)}_{j|\operatorname{Pa}^*_j} || f^{(e)}_{j|\operatorname{Pa}_j}) \right] \right\}
\end{aligned} \tag{27}$$

if Equation (27) is zero, by strictly positive assumption, we have $\mathrm{D}_{\mathrm{KL}}(p^{(a)}_{j|\operatorname{Pa}^*_j} || f^{(e)}_{j|\operatorname{Pa}_j}) = \mathrm{D}_{\mathrm{KL}}(p^{(b)}_{j|\operatorname{Pa}^*_j} || f^{(e)}_{j|\operatorname{Pa}_j}) = 0$,which implies $p^{(b)}_{j|\operatorname{Pa}^*_j} = p^{(a)}_{j|\operatorname{Pa}^*_j}$, contradiction. Thus, Equation (27) is positive.

Now we assume that each estimated cluster $\hat{\mathcal{E}}_c$ ($c \in \{1, \ldots, N_{\mathrm{c}}\}, N_{\mathrm{c}} \geq K$) contains samples from same domains. We rearrange the index of clusters such that samples in the k-th cluster are from the k-th domain for $k \in \{1, \ldots, K\}$. Then Equation (26) has lower bound

$$\begin{aligned}
&\inf_{\phi} \sum_{\ell=1}^{K} \pi^*_{\ell} \mathop{\mathbb{E}}_{x\sim p^{(\ell)}} \mathrm{D}_{\mathrm{KL}}(p^{(\ell)} || f^{(\ell)}) \\
\geq &(\min_{\ell} \pi^*_{\ell}) \inf_{\phi} \sum_{\ell=1}^{K} \mathrm{D}_{\mathrm{KL}}(p^{(\ell)} || f^{(\ell)})
\end{aligned} \tag{28}$$

Equation (28) is positive if and only if

$$\eta(\mathcal{G}, \mathcal{I}) \coloneqq \inf_{\phi} \sum_{\ell=1}^{K} \mathrm{D}_{\mathrm{KL}}(p^{(\ell)} || f^{(\ell)}) \tag{29}$$

is positive. $\mathcal{G}, \mathcal{I}, \phi$ are used by model $f^{(k)}$ in Equation (3).

Recall that the ground truth distribution set is $\{p^{(I)}\}_{I\in\mathcal{I}^*}$ and $|\mathcal{I}^*| = K$, which means there are K domains and each domain has its own distribution and its own intervention target in $\mathcal{I}$. we have assigned each domain a number in $[K] \coloneqq \{1, 2, \ldots, K\}$. For example, for $I \neq J \in \mathcal{I}^*$ we assign $k, \ell \in [K]$ respectively. In Equation (28) we rearranged the recovered indexes so that it is consistent with assigned indexes, and dropped duplicated clusters to make sure each domain has exactly one corresponding cluster. **These operation is for simplifying the analysis.** Now for each domain, the data is ready, the left is to estimate its intervention target (by finding its elements) and distribution density (approximated by training a neural network).

Now we are going to show: the above score would be positive if some estimated augmented graph $\mathcal{G}^{\tilde{\mathcal{I}}_J}$ do not have the same skeleton or v-structure as the ground truth graph. If this could be done, we could conclude that the learned $\mathcal{G}$ and $\mathcal{I}$ is $\mathcal{I}$-Markov Equivalent to the grounded in the sense $M_{\mathcal{I}}(\mathcal{G}) = M_{\mathcal{I}^*}(\mathcal{G}^*)$ according to proposition 3.14 in Yang et al.'s work (2018).

Equation (28) builds a bridge to the score of previous Brouillard et al.'s work. However, in general, their proof cannot be directly applied because some gaps occur in the case where $\emptyset \notin \mathcal{I}$. In the following part of this proof, our contribution is to show Assumption 5.9 can fill these gaps. We use **Brouillard et al.'s work** to refer to Brouillard et al.(2020)'s result.

Consider any $I \neq J$ in $\mathcal{I}^*$, and the corresponding correct J-observation target is $\tilde{I}^*_J$, which is estimated by $\tilde{I}_J$. Suppose any $j \in [p]$ such that $j \in \tilde{I}^*_J$ but $j \notin \tilde{I}_J$. Since node j and $\zeta_{\tilde{I}^*_J}$ are never d-separated in ground truth $\mathcal{G}^{*\tilde{\mathcal{I}}^*_J}$, so we have:

$$j \not\perp\!\!\!\perp_{\mathcal{G}^{*\tilde{\mathcal{I}}^*_J}} \zeta_{\tilde{I}^*_J} | \mathrm{Pa}_j \cup \zeta_{-\tilde{I}^*_J} \tag{30}$$

(recall that Pa_j means parents set of node j in graph $\mathcal{G}$, not in ground truth graph $\mathcal{G}^*$)

The following part of the proof involves a notion called $\mathcal{Z}(j, A)$ used in **Brouillard et al.'s work**.

$$\mathcal{Z}(j, A) \coloneqq \{(p^{(1)}, p^{(1)}) | p^{(1)}_{j|A} = p^{(2)}_{j|A} \text{ and } p^{(1)}, p^{(1)} > 0\} \tag{31}$$

By Assumption 5.9 and second property in Definition 5.5, we have

$$\left[\tilde{p}^{(\tilde{I}_J)}_J\right]_{j|\mathrm{Pa}_j} \neq \left[\tilde{p}^{(\emptyset)}_J\right]_{j|\mathrm{Pa}_j} \tag{32}$$

which means $p^{(I)}_{j|\mathrm{Pa}_j} \neq p^{(J)}_{j|\mathrm{Pa}_j}$, so $(p^{(I)}, p^{(J)}) \notin \mathcal{Z}(j, \mathrm{Pa}_j)$. On the other hand, recall the model design in Equation (3) and Equation (4), suppose the index for I and J is k and ℓ respectively, then $j \notin \tilde{I}_J$, therefore, $f^{(I)}_{j|\mathrm{Pa}_j} = f^{(J)}_{j|\mathrm{Pa}_j}$ and $(f^{(I)}, f^{(J)}) \in \mathcal{Z}(j, \mathrm{Pa}_j)$. So we have

$$\begin{aligned}
\eta(\mathcal{G}, \mathcal{I}) &= \inf_{\phi} \sum_{\ell=1}^{K} \mathrm{D}_{\mathrm{KL}}(p^{(\ell)} || f^{(\ell)}) \\
&\geq \inf_{\phi} \mathrm{D}_{\mathrm{KL}}(p^{(I)} || f^{(I)}) + \mathrm{D}_{\mathrm{KL}}(p^{(J)} || f^{(J)}) \\
&\geq \inf_{(f_1, f_2) \in \mathcal{Z}(j, \mathrm{Pa}_j)} \mathrm{D}_{\mathrm{KL}}(p^{(I)} || f_1) + \mathrm{D}_{\mathrm{KL}}(p^{(J)} || f_2) \\
&> 0
\end{aligned} \tag{33}$$

The last inequality holds by **Lemma 16** in **Brouillard et al.'s work**.

In **Case 0.1** of proof for Theorem 2 in Brouillard et al.'s work, they have shown that if $\eta(\mathcal{G}, \mathcal{I}) > 0$, then $\lambda + \lambda_M$ is sufficient small would implies Equation (26) is positive, which in the sense

$$\lambda + \lambda_M < \min_{(\mathcal{G}, \mathcal{I}) \in \mathbb{S}} \frac{\eta(\mathcal{G}, \mathcal{I})}{-\min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*|\}} \tag{34}$$

where $\mathbb{S} := \{(\mathcal{G}, \mathcal{I}) | \min\{|\mathcal{G}| - |\mathcal{G}^*|, |\mathcal{I}| - |\mathcal{I}^*| < 0\}$.

Thus we have $\tilde{I}^*_J \subset \tilde{I}_J$ and therefore $|\tilde{I}^*_J| \leq |\tilde{I}_J|$. Penalty on $|\tilde{I}_J|$ would push $|\tilde{I}^*_J| = |\tilde{I}_J|$ (there is also an inequality like Equation (34), see **Case 0.2** in their paper), so $\tilde{I}^*_J$ can be learned correctly. The above reasoning is adjusted from **Case 0.1** and **Case 0.2** of proof for Theorem 2 in Brouillard et al.'s work. From now on we assume we found $\mathcal{I}^*$.

Consider any $C \subset V$, and $i \neq j$ in $V \setminus C$, that i is d-connected to j given C in true graph $\mathcal{G}^*$. By Assumption 5.9 and the first condition in Definition 5.5, there exists $I \in \mathcal{I}$, such that $X_i \not\perp\!\!\!\perp_{p^{(I)}} X_j | X_C$. Therefore, in **Case 1 , 3, and 4** of proof for Theorem 1 in Brouillard et al.'s work, one can still find such $p^{(I)} \notin \mathcal{M}(\mathcal{G})$, which implies $\eta(\mathcal{G}, \mathcal{I}^*) > 0$

As shown in **Case 1** of proof for Theorem 1 in Brouillard et al.'s work: if $\eta(\mathcal{G}, \mathcal{I}^*) > 0$, then λ is sufficient small would implies Equation (26) is positive, which in the sense

$$\lambda < \min_{(\mathcal{G}, \mathcal{I}^*) \in \mathbb{G}^+} \frac{\eta(\mathcal{G}, \mathcal{I}^*)}{|\mathcal{G}^*| - |\mathcal{G}|} \tag{35}$$

where $\mathbb{G}^+$ is as same as what is defined in their paper.

In **Case 2**, where nothing needs to be adjusted. And we could assume $|\mathcal{G}| = |\mathcal{G}^*|$ in **Case 3** $\sim$ **6**, in which case, $\eta(\mathcal{G}, \mathcal{I}^*) > 0$ implies Equation (26) is positive.

Consider any $C \subset V$, $i \in V \setminus C$, and $I \in \mathcal{I}$, that i is d-connected to ζ_I given $C \cup \zeta_{-I}$ in true augmented graph $\mathcal{G}^{*\mathcal{I}^*}$. By Assumption 5.9 and second property in Definition 5.5, $[\tilde{p}^{(\tilde{I}_J)}_J]_{j|C} \neq [\tilde{p}^{(\emptyset)}_J]_{j|C}$ i.e. $p^{(I)}_{j|C} \neq p^{(J)}_{j|C}$ Thus, **Case 5, 6** of proof for Theorem 1 in Brouillard et al.'s work one can still find such $(p^{(I)}, p^{(J)}) \notin \mathcal{Z}(j, \mathrm{Pa}_j)$, which implies $\eta(\mathcal{G}, \mathcal{I}^*) > 0$

Now, we have shown that with the given condition, for any $J \in \mathcal{I}$, we can learn each J-observation augmented graph's skeleton and v-structure. By Theorem 3.14 in Yang et al.(2018)'s paper, we learned an $\mathcal{I}$-MEC to the true augmented graph.

□

D. Detailed Implementation

Hyper-parameter setting In experiments, each method requires multiple hyper-parameters. We highlight how those hyper-parameters are set in Table 8. Both synthetic and real-world datasets are about time series, we control the maximal time lag to be 1, which means all methods consider the relation between x_{t-1} and x_t; and ignore relations like x_{t-2} and x_t. Some methods rely on Conditional Independence tests (i.e. CI Test), each test gives a p-value to compare with a pre-specified value α, called confidence level. α is controlled to be 0.01. Some methods require coefficients for penalty terms. The magnitude of those coefficients depends on the specific implementations of specific methods. Therefore, different methods have different magnitudes for penalty-term coefficients.

Table 8. Hyper-parameter setting

HYPER-PARAMETER	DESCRIPTION	INVOLVED METHODS	VALUE
TIME LAG	TIME LAG RELATION TO CONSIDER	ALL	1
ALPHA	CONFIDENCE LEVEL FOR CI TESTS	THOSE WHO USE CI TEST	0.01
HIDDEN_DIM	NEURAL NETWORK ARCHITECTURE	DCDI, LIN	5
N_HIDDEN_LYR	NEURAL NETWORK ARCHITECTURE	DCDI, LIN	1
LR	LEARNING RATE	DCDI, LIN	0.01
LMD	COEFFICIENT FOR PENALTY TERMS	DYNOTEARS	[1E-1, 1E-3]
PNT	COEFFICIENT FOR PENALTY TERMS	DCDI	[1E-2, 1E-4]
PNT	COEFFICIENT FOR PENALTY TERMS	LIN	{1E-08, 1E-16}
N_c	NUMBER OF CLUSTERS TO CONSIDER	LIN	{2, 3, 4, 5}

Tuning For baseline methods, if there exist multiple choices for hyper-parameters, we take the result with the best F1 score in the test set. For our LIN method, we take the result with the best criterion as stated in Equation (11). This criterion is calculated in hold-out data in the training set.

Interpreting PAGs as DAGs Some baseline methods, like CD-NOD, PCMCI, and SVAR-FCI, would not guarantee DAGs, instead, they report Partial Ancestral Graphs (PAGs). Therefore, inspired by Pamfil et. al. (2020), we use following rules in evaluation process:

- For usual directed edges $\longrightarrow$: we treat them as regular edges in DAGs.
- For bi-directed edges $\longleftrightarrow$: this type of edge means there is no causal relation between two variables (although they are correlated), so we drop them before evaluation.
- For other types of edges with ambiguous, like $\circ\!\!\longrightarrow$ and $\circ\!\!-\!\!\circ$: only check whether skeleton is correct.

with these rules, we got results in evaluation, which prefer to believe those ambiguous edges are correct.

Details about Generating Process driven by Brownian Motions In this paper, we generate two groups of synthetic data by the dynamic behind those latent interventions. The first group is those whose interventions are controlled by a Brownian motion, as illustrated in Figure 3. The left part is one-dimension Brownian motion, the horizon axis is time t, and the vertical axis can be seen as the position of the system concerned in its environment; Different colors represent domains with different interventional distributions. For example, when time $t = 2$, the system is in the red area, its data generating process follows the interventional distribution associated with this red area $P_{\text{red}}(x_t|x_{t-1})$; and when time $t = 6$, it follows the interventional distribution associated with blue area $P_{\text{blue}}(x_t|x_{t-1}) \neq P_{\text{red}}(x_t|x_{t-1})$. The sample partition can be represented in time-order like

$$\{\text{red}, \text{red}, \ldots, \text{blue}, \text{red}, \ldots, \text{blue}, \text{blue}\}$$

or we can use domain index $z_t \in \{0, 1\}$ to represent the interventional distribution from which sample x_t is generated. If we use 0 for*red*, and use 1 for *blue*, then domain indexes are:

$$\{0, 0, \ldots, 1, 0, \ldots, 1, 1\}$$

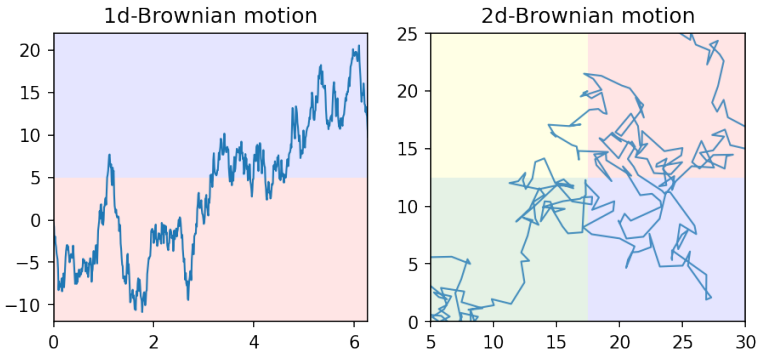

Figure 3. examples for Brownian motion

Similarly, The right part is two-dimension Brownian motion in the x-y plain which is parameterized by time t. In this case, there are four possible interventional distributions that may affect the system we concern: $P_{\text{red}}(x_t|x_{t-1})$, $P_{\text{blue}}(x_t|x_{t-1})$, $P_{\text{yellow}}(x_t|x_{t-1})$, and $P_{\text{green}}(x_t|x_{t-1})$; and domain index $z_t \in \{0, 1, 2, 3\}$ can be used.

When generating synthetic data, we let each colored area be a *finite* interval with length M in the one-dimensional world; or an $M \times M$ rectangle in the two-dimensional world, repeated in whole space, and set $M = 5$ in practice. Each node's interventional distribution is controlled by a neural network that transforms its parents in the causal graph to distribution parameters. In practice, we use Gaussian distribution, and the parameters mean, and variance are determined by nodes' parents.

Although for synthetic data, we know the exact value of domain index z_t, they are assumed to be unknown as in realistic situations, and would only be used to evaluate the model's ability to recover them. None of our proposed LIN method or other baseline methods can use these domain indexes in training.

Extraction of Real-world Climate dataset We use *ERA5 hourly data on single levels from 1959 to present* in *Copernicus climate data store* the variables are *2m temperature* and *Surface pressure*. The data is sub-sampled to 00:00 on each day from 2001 to 2021. The latitude and longitude for Pacific Walker Data are summarised in Table 9. For each region, we describe a rectangle in the world map, and valuables in that region are averaged so that for each one of the four regions, there is a scalar value at each time point.

Table 9. latitude and longitude

SUB-REGION	NORTH	WEST	EAST	SOUTH	VARIABLE
WPAC	5	140	145	-5	SURFACE PRESSURE
CPAC	5	-145	-140	-5	SURFACE AIR TEMPERATURE
EPAC	5	-95	-90	-5	SURFACE AIR TEMPERATURE
ATL	20	-50	-40	10	SURFACE AIR TEMPERATURE