IDENTIFYING PARTIALLY OBSERVED CAUSAL MODELS FROM HETEROGENEOUS/NONSTATIONARY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating causal structure in the presence of latent variables is an important yet challenging problem. Recent works have shown that distributional constraints, such as rank deficiency constraints of the covariance matrices, can be exploited to recover the underlying causal structure involving latent variables. However, real-world data often exhibit heterogeneity/nonstationarity, which pose challenges to existing methods. In this work, we develop a principled approach for identifying the structure of partially observed linear causal models from heterogeneous/nonstationary data. We first formulate a class of heterogeneous/nonstationary, partially observed linear causal models and prove that their distributional constraints are equivalent to those in the homogeneous case. Building on this, we propose a novel rank test that can efficiently handle heterogeneous/nonstationary data, and further establish identifiability results for recovering the causal structure involving latent variables. We also provide a method to identify which variables exhibit distribution shifts, i.e., whose causal mechanisms vary across domains. Experiments on simulated and real-world data validate our theoretical findings and the effectiveness of our method (code will be available).

1 INTRODUCTION AND RELATED WORK

Discovering causal relations from data is one of the fundamental challenges in many scientific disciplines (Spirtes et al., 2000; Pearl et al., 2000). Traditional causal discovery methods typically assume causal sufficiency, indicating that there are no latent confounders (Spirtes et al., 2000; Spirtes, 2001; Chickering, 2002). However, this assumption may often be violated in practice and ignoring these latent variables can lead to inaccurate causal conclusions. This highlights the importance of causal discovery methods that account for latent confounders.

To address this challenge, early methods such as Fast Causal Inference (FCI) (Spirtes et al., 2000; Zhang, 2008) and its variants (Colombo et al., 2012; Spirtes et al., 2013; Claassen et al., 2013; Akbari et al., 2021) utilize conditional independence tests to identify causal relations among observed variables while accounting for latent confounders. These methods output partial ancestral graphs (PAGs) (Richardson, 1996) over the observed variables, which summarize the equivalence class of causal structures consistent with the data. While FCI does not make any assumption about the latent structure, it often produces less informative outputs, e.g., provides no information about relationships among latent variables. In contrast, recent approaches aim to recover the full causal structure, including latent-to-latent and latent-to-observed relations, by leveraging parametric or graphical assumptions. These approaches are typically based on tetrad or rank constraints (Silva et al., 2003; 2006; Choi et al., 2011; Kummerfeld & Ramsey, 2016; Huang et al., 2022; Dong et al., 2024), higher-order moments (Shimizu et al., 2009; Cai et al., 2019; Salehkaleybar et al., 2020; Xie et al., 2020; Adams et al., 2021; Dai et al., 2022; Chen et al., 2022; Améndola et al., 2023; Wang & Drton, 2023), matrix decompositions (Anandkumar et al., 2013), and score-based search (Ng et al., 2024).

Apart from latent confounders, another challenge in real-world settings is the presence of heterogeneity in the data. Such variation often arises from different types of interventions, ranging from hard interventions (Eberhardt & Scheines, 2007; Hauser & Bühlmann, 2012) to soft interventions (Eberhardt & Scheines, 2007; Yang et al., 2018). To address this, various constraint-based (Huang et al., 2020a), score-based (Hauser & Bühlmann, 2012; Squires et al., 2020; Brouillard et al., 2020), and hybrid (Wang et al., 2017; Yang et al., 2018) methods, as well as other general frameworks (Mooij et al., 2020), have been proposed to infer causal structure from interventional or heterogeneous data.

To handle both latent confounders and heterogeneous/nonstationary data, Magliacane et al. (2016); Kocaoglu et al. (2019) proposed constraint-based methods that rely on conditional independence tests to recover ancestral structures over the observed variables, similar in spirit to FCI. As previously discussed, these outputs may often be uninformative when the goal is to understand the relationships among latent variables. In contrast, a related line of work, causal representation learning (Schölkopf et al., 2021), aims to infer both the latent causal variables and the causal structure among them. While these methods also leverage interventional or heterogeneous data (Hyvärinen et al., 2023; Ahuja et al., 2023; Squires et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Jin & Syrgkanis, 2023; Zhang et al., 2024; Varıcı et al., 2024a;b; Bing et al., 2024; Ng et al., 2025), they typically make certain assumptions: (1) no causal edges exist among observed variables or from observed to latent variables, and (2) the generative process from latent variables is deterministic (except several works including Khemakhem et al. (2020); Lachapelle et al. (2024); Fu et al. (2025)) and invariant across domains. Here, we consider a more general setting that relaxes these assumptions. Further discussions of the related works are provided in Appendix D.

In this work, we consider a general setting that can handle latent variables and heterogeneous/nonstationary data, allowing all variables, including both observed and latent ones, to be flexibly related. Our contributions are summarized as follows:

- We formulate a class of heterogeneous/nonstationary, partially observed linear causal models, i.e., Nonstationary POLCMs (Definition 1), to properly handle nonstationary data. It allows changing model parameters both within and across domains while preserves structure identifiability.
- We prove that, despite the existence of nonstationarity, the conditional covariance set generated by Nonstationary POLCMs are equivalent to the covariance set in the homogeneous case (Theorem 1). This implies all the constraints on conditional distribution imposed by structure are equivalent to those in the homogeneous case (Corollary 1), and thus the possibility of those constraint-based homogeneous causal discovery methods to be upgraded to handle the nonstationary scenario.
- To make use of equality constraints for structure identifiability, we establish the relation between rank of conditional covariance and t-separations for Nonstationary POLCMS (Theorem 2); notably, it takes the relation between vanishing partial correlation and d-separations as a special case. To properly control statistical errors with finite data, we further propose a novel statistical test to examine the rank of conditional covariance in the nonstationary scenario (Theorem 3).
- We propose a novel method, Latent variable Causal Discovery from heterogeneous/NOnstationary Data (LCD-NOD), to identify the structure of Nonstationary POLCMs. The first phase serves as a general augmentation of current equality constraint-based methods, e.g., PC (Spirtes et al., 2000), FOFC (Silva et al., 2003; Kummerfeld & Ramsey, 2016), and RLCD (Dong et al., 2024), to handle nonstationary data, while the second phase further identifies which variables are directly influenced by the nonstationarity. Extensive experiments validate the proposed rank test and LCD-NOD using both synthetic and real-life data.

2 PRELIMINARIES

2.1 PROBLEM SETTING

To better handle both nonstationarity and latent variables, we assume that data is generated by Nonstationary Partially Observed Linear Causal Models (nonstationary POLCMs), defined as follows.

Definition 1 (Nonstationary POLCMs). *Let $\mathcal{G}$ be a DAG with variable set $\mathbf{V} = \mathbf{X} \cup \mathbf{L} = \{\mathsf{X}_i\}^n \cup \{\mathsf{L}_i\}^m$ that contains n observed and m latent variables. Each variable $\mathsf{V}_i \in \mathbf{V}$ is generated following*

$$\mathsf{V}_i = \sum\nolimits_{\mathsf{V}_j \in Pa_{\mathcal{G}}(\mathsf{V}_i)} h_{j,i}(\mathsf{T}, \delta_{j,i})\mathsf{V}_j + g_i(\mathsf{T}, \epsilon_i), \tag{1}$$

where $Pa_{\mathcal{G}}(\mathsf{V}_i)$ denotes the parent set of V_i, $h_{j,i}(\mathsf{T}, \delta_{j,i})$ denotes the edge coefficient from V_j to V_i, and $\delta_{j,i}$ and ϵ_i are independent noise terms.

Remark 1. *In Definition 1, T can be understood as the domain index. The coefficient for edge $\mathsf{V}_j \to \mathsf{V}_i$ is $h_{j,i}$, which is a deterministic function of T and $\delta_{j,i}$. The additive noise term g_i is also a determinstic function of T and ϵ_i. Therefore, in Equation (1), two kinds of nonstationarity can be modeled. 1. nonstationarity across domains, as both $h_{j,i}$ and g_i are functions of domain index. 2. Nonstationarity within domain, as edge coefficients $h_{j,i}$ is also a function of independent noise term $\delta_{j,i}$. It is possible to model these two kinds of nonstationarity in a more complex functional fashion,*

Table 1: Graphical notations used throughout this paper.

Pa: Parents	$\mathbf{V}$: Variables	V: Variable	T: Domain / Time index
Ch: Children	$\mathbf{L}$: Latent variables	L: Latent variable	$\mathcal{G}$: Ground truth structure
PCh: Pure children	$\mathbf{X}$: Observed variables	X: Observed variable	$\mathcal{G}^{\text{aug}}$: Structure involving T

and yet one major goal of this work is to show that the constraints are equivalent to those in the homogeneous case. If the functional form can be arbitrary, then Theorem 1 might not hold anymore and we conjecture that any further relaxation of functional form would induce failure of Theorem 1.

Given i.i.d. samples of observed variables $\mathbf{X}$ generated by Definition 1, our objective is to identify the causal structure of the underlying nonstationary causal model. More specifically, we aim to identify the causal structure among all the variables, including both observed and latent ones. Further, the generating process for some variables might be stationary. Thus we are also interested in identifying the stationary and nonstationary variable set, or equivalently, identifying $\mathcal{G}^{\text{aug}}$, where $\mathcal{G}^{\text{aug}}$ is the augmented graph of $\mathcal{G}$ such that $\mathcal{G}^{\text{aug}}$ contains one additional node T and $\mathsf{T} \to \mathsf{V}_i$ if and only if $h_{:,i}$ and g_i can change with different values of T.

2.2 Notations and Paper Organization

In this paper, we use V to denote random variable and $\mathbf{V}$ a set of random variables. We use T to refer to the random variable that represents domain index, and t the observed value. A summary of commonly used notations can be found in Table 1.

The rest of the paper is organized as follows. In Section 3.1, we briefly introduce the constraints in the homogeneous setting for structure identification and show that they normally fail to hold in the nonstationary scenario. In Section 3.2, we formally show the equivalence relation between constraints on the covariance in the homogeneous case and the constraints on the conditional covariance in the nonstationary case. In Section 3.3, we focus on rank constraints and characterize their graphical implications for structure identification, with a novel rank test of conditional covariance proposed in Section 3.4. In Section 4, we propose the Latent variable Causal Discovery from heterogeneous/NOnstationary Data (LCD-NOD) algorithm, and empirically validate LCD-NOD in Section 5.

3 Distributional Information for Structure Identification

3.1 Constraints in the Homogeneous Case

In this section, we first revisit the constraints in homogeneous Partially Observed Linear Causal Models (POLCMs), defined as follows,

Definition 2 (Homogeneous POLCMs)**.** *Let $\mathcal{G}$ be a DAG with variable set $\mathbf{V} = \mathbf{X} \cup \mathbf{L} = \{\mathsf{X}_i\}^n \cup \{\mathsf{L}_i\}^m$ that contains n observed and m latent variables. Each variable $\mathsf{V}_i \in \mathbf{V}$ is generated following*

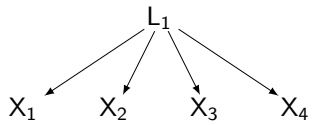
$$\mathsf{V}_i = \sum\nolimits_{\mathsf{V}_j \in Pa_{\mathcal{G}}(\mathsf{V}_i)} f_{j,i}\mathsf{V}_j + \epsilon_i, \tag{2}$$

where $Pa_{\mathcal{G}}(\mathsf{V}_i)$ denotes the parent set of V_i, $f_{j,i}$ denotes the edge coefficient from V_j to V_i, and ϵ_i are independent noise terms.

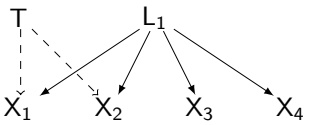
For a model in Definition 2, its structure $\mathcal{G}$ imposes various constraints on the generated population covariance matrices, regardless of the parameter values ($(f_{j,i})$ and variance of ϵ_i). These constraints fall into two categories: equality and inequality constraints. For structure identification, equality constraints are most informative. They include, for example, conditional independence (i.e., vanishing partial correlation) constraints (Spirtes et al., 2000), which suffice to identify the whole structure up to the Markov Equivalence Class (MEC) when there are no latent variables. To handle latent variables, more equality constraints have been discovered and exploited, including rank constraints (i.e., vanishing determinant) (Sullivant et al., 2010), Verma constraints (Verma & Pearl, 1991), among others. An overview can be found in Drton (2018).

Given that these equality constraints contain crucial information for structure identifiability, a question naturally arises: Can these constraints in the homogeneous case be directly extended to the nonstationary setting? Unfortunately, they do not generally carry over, as illustrated by the following example.

Example 1. *In Figure 1, the model in (a) follows Definition 2 and the model in (b) follows Definition 1, but they share the same DAG structure among $\mathbf{V}$. However, the equality constraints in*



(a) $\mathcal{G}$ under the homogeneous POLCMs setting, where the edge coefficients and the variance of the noise terms are fixed parameters.



(b) $\mathcal{G}^{\text{aug}}$ under nonstationary setting where the dashed arrows from T to X_1 and X_2 means the edge coefficients from L_1 to X_1 and X_2 change across T.

Figure 1: Illustrative example to show that, with the same structure but in the presence of nonstationarity, the same equality constraint does not hold anymore. Specifically, in (a) $\sigma_{\mathsf{X}_1,\mathsf{X}_2}\sigma_{\mathsf{X}_3,\mathsf{X}_4} = \sigma_{\mathsf{X}_1,\mathsf{X}_3}\sigma_{\mathsf{X}_2,\mathsf{X}_4}$ holds regardless of the choice of parameters, while in (b) it does not.

(a) may not hold any more in (b). For example, the classical tetrad constraint (a kind of equality constraint) implies $\sigma_{\mathsf{X}_1,\mathsf{X}_2}\sigma_{\mathsf{X}_3,\mathsf{X}_4} = \sigma_{\mathsf{X}_1,\mathsf{X}_3}\sigma_{\mathsf{X}_2,\mathsf{X}_4}$ in (a) regardless of the choice of parameter in (a), as $\frac{\sigma_{\mathsf{X}_1,\mathsf{X}_2}\sigma_{\mathsf{X}_3,\mathsf{X}_4}}{\sigma_{\mathsf{X}_1,\mathsf{X}_3}\sigma_{\mathsf{X}_2,\mathsf{X}_4}} = \frac{\mathbb{E}[f_{\mathsf{L}_1,\mathsf{X}_1} f_{\mathsf{L}_1,\mathsf{X}_2}]\mathbb{E}[f_{\mathsf{L}_1,\mathsf{X}_3} f_{\mathsf{L}_1,\mathsf{X}_4}]}{\mathbb{E}[f_{\mathsf{L}_1,\mathsf{X}_1} f_{\mathsf{L}_1,\mathsf{X}_3}]\mathbb{E}[f_{\mathsf{L}_1,\mathsf{X}_2} f_{\mathsf{L}_1,\mathsf{X}_4}]} = 1$ always holds due to that $(f_{j,i})$ are constants. However, in (b) $\frac{\sigma_{\mathsf{X}_1,\mathsf{X}_2}\sigma_{\mathsf{X}_3,\mathsf{X}_4}}{\sigma_{\mathsf{X}_1,\mathsf{X}_3}\sigma_{\mathsf{X}_2,\mathsf{X}_4}} = \frac{\mathbb{E}_{\mathsf{T},\delta,\epsilon}[h_{\mathsf{L}_1,\mathsf{X}_1} h_{\mathsf{L}_1,\mathsf{X}_2}]\mathbb{E}_{\mathsf{T},\delta,\epsilon}[h_{\mathsf{L}_1,\mathsf{X}_3} h_{\mathsf{L}_1,\mathsf{X}_4}]}{\mathbb{E}_{\mathsf{T},\delta,\epsilon}[h_{\mathsf{L}_1,\mathsf{X}_1} h_{\mathsf{L}_1,\mathsf{X}_3}]\mathbb{E}_{\mathsf{T},\delta,\epsilon}[h_{\mathsf{L}_1,\mathsf{X}_2} h_{\mathsf{L}_1,\mathsf{X}_4}]}$ where the expectation is taken over $\mathsf{T}, \delta, \epsilon$ and both $h_{\mathsf{L}_1,\mathsf{X}_1}$ and $h_{\mathsf{L}_1,\mathsf{X}_2}$ are functions of T, δ, and thus the equality constraint does not generally hold.

By the above example, we know that the equality constraints does not readily transfer to the nonstationary case. Thus we need to characterize the equality constraints in the nonstationary case for structure identification, detailed as follows.

3.2 Constraints Implied by Structure Under Nonstationarity

In this section, we aim to characterize useful constraints in the nonstationary scenario for structure identification. To this end, we need to first define the observational covariance set under POLCMs, as $\Theta(\mathcal{G})$, and the observational conditional covariance set under Nonstationary POLCMs, as $\Phi(\mathcal{G})$, in Definition 3 and Definition 4, respectively. The reason why we care about $\Theta(\mathcal{G})$ and $\Phi(\mathcal{G})$ is as follows. A constraint imposed by $\mathcal{G}$ on the covariance matrix is nothing but some relations among the entries of the covariance matrix that always holds regardless of the choice of the parameters; Therefore, constraints imposed by $\mathcal{G}$ under POLCMs and Nonstationary POLCMs are just properties of $\Theta(\mathcal{G})$ and $\Phi(\mathcal{G})$ respectively.

Definition 3 (Observational covariance set under POLCMs)**.** *Let $F = (f_{j,i})$ and Ω be the covariance matrix for $\{\epsilon_i\}^{n+m}$. We define the observational covariance set of $\mathcal{G}$ under POLCMs (Definition 2) as:*

$$\Theta(\mathcal{G}) := \{\Theta : \Theta = ((I - F^T)^{-1}\Omega(I - F^T)^{-T})_{[:n,:n]}, \tag{3}$$

$$\text{for any } (F, \Omega) \text{ s.t. } \Omega \in \text{diag}^+ \text{and } \text{supp}(F) \subseteq \text{supp}(F_{\mathcal{G}})\}. \tag{4}$$

Remark 2. *The meaning of $\Theta(\mathcal{G})$ is just the set of all possible covariance matrices over $\{\mathsf{X}_i\}^n$, by taking the same causal structure $\mathcal{G}$ and different model parameters (F, Ω). Note that if we assume all ϵ_i in Definition 2 follow Gaussian distributions, then $\{\mathsf{X}_i\}^n$ are jointly Gaussian. In this case, only the second-order information matters, and thus all the distributional information that can be used for structure identification are just all the properties of $\Theta(\mathcal{G})$.*

Definition 4 (Observational conditional covariance set under Nonstationary POLCMs)**.** *Let $H(\mathsf{T}) = (h_{j,i}(\mathsf{T}, \delta_{j,i}))$ and $g(\mathsf{T}) = (g_i(\mathsf{T}, \epsilon_i))$. The observational conditional covariance set of $\mathcal{G}$ under Nonstationary POLCMs (Definition 1) is defined as:*

$$\Phi(\mathcal{G}) := \{\Phi : \Phi = \mathbb{E}_{p(\mathbf{V}|\mathsf{T})}[(I - H(\mathsf{T})^T)^{-1} g(\mathsf{T}) g(\mathsf{T})^T (I - H(\mathsf{T})^T)^{-T}]_{[:n,:n]}, \tag{5}$$

$$\text{for any } (H(\mathsf{T}), g(\mathsf{T})) \text{ s.t. } \text{supp}(H(\mathsf{T})) \subseteq \text{supp}(H_{\mathcal{G}})\}. \tag{6}$$

Remark 3. *Similar to $\Theta(\mathcal{G})$, $\Phi(\mathcal{G})$ is the set of all possible conditional covariance matrices over $\{\mathsf{X}_i\}^n$, by taking the same $\mathcal{G}$ and different model parameters (H, g). We note that the definition of $\Phi(\mathcal{G})$ involve conditioning on T, but for any specific value of T, $\Phi(\mathcal{G})$ keeps the same, and thus we omit the notion T in $\Phi(\mathcal{G})$.*

Given the definitions of $\Theta(\mathcal{G})$ and $\Phi(\mathcal{G})$, we now introduce the main theoretical result of this section:

Theorem 1 (Equivalence Relation between $\Theta(\mathcal{G})$ under POLCMs and $\Phi(\mathcal{G})$ under Nonstationary POLCMs)**.** *Given a DAG $\mathcal{G}$, let $\Theta(\mathcal{G})$ be the observational covariance set of $\mathcal{G}$ under POLCMS, as in Definition 3, and let $\Phi(\mathcal{G})$ be the observational conditional covariance set of $\mathcal{G}$ under Nonstationary POLCMs, as in Definition 4. We have $\Theta(\mathcal{G}) = \Phi(\mathcal{G})$.*

Theorem 1 says that, given the same structure $\mathcal{G}$, despite the presence of both inter-domain and intra-domain nonstationarity in Nonstationary POLCMs, the possible covariance set in the homogeneous case is exactly the same as the possible conditional covariance set in the nonstationary case. As constraints are just properties of the covariance / conditional covariance set, it can be inferred from Theorem 1 that all the constraints on $\Theta(\mathcal{G})$ are the same as those on $\Phi(\mathcal{G})$, formalized in Corollary 1.

Corollary 1 (Equality and Inequality Constraints Under Nonstationary)**.** *If $\mathcal{G}$ implies an equality or inequality constraint on $\Theta(\mathcal{G})$, then $\mathcal{G}$ also implies the same constraint on $\Phi(\mathcal{G})$, and vice versa.*

Theorem 1 and Corollary 1 are significant, as they prove the existence of structure identifiability of Nonstationary POLCMs. The key takeaways are two folds.

(i): All the constraints implied by $\mathcal{G}$ on $\Theta(\mathcal{G})$ remains on $\Phi(\theta)$; Thus all the equality-constraint-based homogeneous causal discovery methods, e.g., PC (vanishing partial correlation) (Spirtes et al., 2000), FOFC (Tetrad) (Silva et al., 2003; Kummerfeld & Ramsey, 2016), and RLCD (rank constraints) (Dong et al., 2024), can be upgraded to handle the nonstationary scenario (detailed in Section 4).

(ii): If we do not restrict the function forms of h, g and distributions of δ, ϵ in the Nonstationary POLCMs, we cannot determine whether the distribution over $\mathbf{X}$ conditioned on T is jointly gaussian or not. In this case, only the second-order information can be used and thus the graphical information from the constraints on $\Phi(\mathcal{G})$ is complete - those constraints on $\Phi(\mathcal{G})$ are all the information that we can use for structure identification from the observed conditional distribution $p(\mathbf{X}|\mathsf{T})$.

3.3 Rank Constraint under Nonstationarity and its Graphical Implication

In Section 3.2, we have shown the equality constraints in the homogeneous case can be transferred to the constraints in the nonstationary case. In this section, we focus on a specific class of equality constraints, rank deficiency constraints (Sullivant et al., 2010), to establish its relation with trek-separations (definition of trek in Definition 7 while definition of trek-separation in Definition 5) and thus the structure identifiability for Nonstationary POLCMs.

The reason why we focus on rank constraints are as follows. (i) Rank constraints imply t-separations (Sullivant et al., 2010), and thus contain useful graphical information about latent variables. By making use of rank constraints, we can employ RLCD algorithm (Dong et al., 2024) to identify latent variable structures where all variables, including both observed and latent ones, can be very flexibly related. Thus, when rank constraints are properly characterized in the nonstationary setting, we can elegantly upgrade RLCD to identify the structure of Nonstationary POLCMs. (ii) The set of all rank constraints is itself a very large subset of the set of all equality constraints. In fact, rank constraints take both vanishing partial correlation constraints and Tetrad constraints as special cases (Sullivant et al., 2010; Dong et al., 2024). That is to say, when rank constraints are properly characterized, we can translate the vanishing partial correlation constraints and Tetrad constraints into rank constraints, and thus readily upgrade the PC algorithm (Spirtes et al., 2000) (based on vanishing partial correlation) and algorithms for One Factor Models (Silva et al., 2003) (based on Tetrad).

Definition 5 (T-separation (Sullivant et al., 2010))**.** *Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C_A}$, and $\mathbf{C_B}$ be four subsets of $\mathbf{V}_\mathcal{G}$ in graph $\mathcal{G}$ (not necessarily disjoint). ($\mathbf{C_A}$,$\mathbf{C_B}$) t-separates $\mathbf{A}$ from $\mathbf{B}$ if for every trek (P_1,P_2) from a vertex in $\mathbf{A}$ to a vertex in $\mathbf{B}$, either P_1 contains a vertex in $\mathbf{C_A}$ or P_2 contains a vertex in $\mathbf{C_B}$.*

Below we introduce the main theoretical result in this section: rank constraints for Nonstationary POLCMs and its graphical implication of t-separations. This result, to our best knowledge, is the first extension of the characterization of rank and t-separations to the nonstationary setting.

Theorem 2 (Rank and T-separation for Nonstationary POLCMs)**.** *Given two sets of variables $\mathbf{A}$ and $\mathbf{B}$ generated by Nonstationary POLCMs with DAG $\mathcal{G}$ and assume rank faithfulness Dong et al. (2024). We have:*

$$\textit{rank}(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) = \min\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A}, \mathbf{C_B}) \textit{ t-separates } \mathbf{A} \textit{ from } \mathbf{B} \textit{ in } \mathcal{G}\}, \tag{7}$$

where $\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}$ is the cross-covariance over $\mathbf{A}$ and $\mathbf{B}$ conditional on T.

3.4 Valid Rank Test for Rank of Conditional Covariance

In Section 3.3, we formally characterize the relation between rank and t-separations for Nonstationary POLCMs, as in Theorem 2. Note that Theorem 2 requires population conditional covariance, and yet in real-life applications we only have access to finite samples and thus the empirical counterpart.

To properly control statistical errors, we need a valid statistical test to test the rank of the underlying population conditional covariance, by making use of empirical observations.

A straight forward way is just to employ the classical rank test (Anderson, 1984; Camba-Méndez & Kapetanios, 2009), by using data conditioned on a value of T. E.g., assume we are given i.i.d. observations of $\mathbf{X}$ and T, where $\mathsf{T} \in \{1, 2\}$. We can just pick the data points such that $\mathsf{T} = 1$, and use these samples to do classical rank tests. The problem of this straight forward method is that, when the number of domains is large, we only make use of a very small portion of data. E.g., assume that $\mathsf{T} \in \{1, ..., 100\}$ with uniform distribution and we are given 1000 data points. This straight forward method has to condition on a value of T and thus only makes use of 10 data points for the test. Thus, a valid test that can utilize all the data points simultaneously would be favorable.

To this end, we propose a novel statistical test based on likelihood ratio statistics to test the rank of conditional covariance, formalized in Theorem 3.

Theorem 3 (Likelihood Ratio Statistics for Rank of Conditional Covariance)**.** *Given two sets of variables $\mathbf{A}$ and $\mathbf{B}$ with $|\mathbf{A}| = P, |\mathbf{B}| = Q$ and $\mathbf{A} \cup \mathbf{B} = \mathbf{X}$ are jointly gaussian given T. Assume that the null hypothesis $\mathcal{H}_0$ is rank$(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) \leq k$, the likelihood ratio statistics is as follows:*

$$\Lambda(k) = \sum_{t \in supp(\mathsf{T})} - \left(N_t - \frac{P + Q + 1}{2} \right) \ln(\Pi_{i=k+1}^{\min(P,Q)}(1 - r_{t,i}^2)), \tag{8}$$

where N_t is the number of data points such that $\mathsf{T} = t$, and $r_{t,i}$ is the i-th canonical correlation between $\mathbf{A}$ and $\mathbf{B}$ conditioned on $\mathsf{T} = t$. We have that $\Lambda(k)$ converges in distribution to χ^2_{df}, with degree of freedom $df = \sum_{t \in supp(\mathsf{T})}(P - k)(Q - k)$.

To perform the test, we just calculate the test statistics $\Lambda(k)$ and plug it in to the corresponding chi-square distribution to get the p-value. As shown in Theorem 3, the proposed test statistics $\Lambda(k)$ is a likelihood ratio statistics based on $p(\mathbf{X}, \mathsf{T})$ instead of $p(\mathbf{X}|\mathsf{T})$, and thus it makes use of all the data points instead of those conditioned on a specific value of T. Further, as a likelihood ratio test, its asymptotic optimality in terms of power can normally be guaranteed under regularity conditions (Van der Vaart, 2000; Lehmann et al., 1986). This is also empirically validated in Section 5.2 where the proposed test properly controls Type-I and has smaller Type-II errors compared to baselines.

4 Latent Variable Causal Discovery from NOnstationary Data

In this section, we present LCD-NOD (Latent variable Causal Discovery from heterogeneous/NOnstationary Data), a two-phase algorithm that first recovers the graph $\mathcal{G}$ from single-domain data, and then identifies changes across domains, represented by the augmented graph $\mathcal{G}^{\text{aug}}$.

4.1 Phase 1: Identification of Structure $\mathcal{G}$

The first phase of LCD-NOD aims to identify the causal structure $\mathcal{G}$ among all the variables $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$. It is designed as a general augmentation of existing equality-constraint-based methods to work in the nonstationary scenario. The key idea is simple and effective: for a given causal discovery algorithm, replace the test of equality constraint on the covariance matrices by the proposed conditional rank test in Theorem 3. Next, we take PC and RLCD as two examples to show how Phase 1 works.

(i) For PC algorithm (Spirtes et al., 2000) (and those based on vanishing partial correlations such as FCI (Spirtes et al., 2000; Zhang, 2008)), replace the test of $\mathsf{V}_1 \perp\!\!\!\perp \mathsf{V}_2 | \mathbf{V_3}$ by the rank test using Theorem 3. Specifically, if we fail to reject rank$(\Sigma_{\{\mathsf{V}_1\} \cup \mathbf{V_3}, \{\mathsf{V}_2\} \cup \mathbf{V_3}|\mathsf{T}}) \leq |\mathbf{V_3}|$, then we fail to reject $\mathsf{V}_1 \perp\!\!\!\perp \mathsf{V}_2 | \mathbf{V_3}$. **(ii)** For rank based methods, e.g., Hier-Rank (Huang et al., 2022) and RLCD (Dong et al., 2024), replace the test of rank$(\Sigma_{\mathbf{V_1}, \mathbf{V_2}}) \leq k$ by the rank test using Theorem 3. Specifically, if we fail to reject rank$(\Sigma_{\mathbf{V_1}, \mathbf{V_2}|\mathsf{T}}) \leq k$, then we fail to reject rank$(\Sigma_{\mathbf{V_1}, \mathbf{V_2}}) \leq k$. Further, for those methods that are based on Tetrad constraints, we can just reformulate the Tetrad constraint into a rank constraint and test it using Theorem 3.

LCD-NOD Phase 1 serves as a general framework in which any existing latent variable causal discovery method based on covariance information (e.g., for linear Gaussian models) can be directly generalized to handle nonstationary data. This is justified by Theorem 1 which shows that in each domain, the induced covariance set remains identical to that of a static model despite nonstationarity.

We conclude introducing LCD-NOD Phase 1 by formally stating its structure identifiability:

Theorem 4 (Structure Identifiability of LCD-NOD Phase 1)**.** *Given an equality-constraint-based causal discovery method $\mathcal{M}$, if $\mathcal{M}$ asymptotically identifies $\mathcal{G}$ up to equivalence class $\mathcal{C}$ under POLCMs and graphical assumption $\mathcal{A}$, then the Phase 1 of LCD-NOD, i.e., the augmented $\mathcal{M}$, asymptotically identifies $\mathcal{G}$ up to $\mathcal{C}$, under Nonstationary POLCMs and $\mathcal{A}$.*

4.2 Phase 2: Identification of Augmented Structure $\mathcal{G}^{\text{AUG}}$

In previous sections, we show that any existing latent variable causal discovery method based on covariance information can be directly applied using each single domain's information. In this section, we go further by leveraging data across different domains to identify where changes happen. For a case study, we consider a specific model, the one-factor model (Silva et al., 2003).

One-factor model captures cases where latent variables are indirectly measured. By introducing non-stationarity, it becomes a special case of the Nonstationary POLCM model (Definition 1) as follows:

Definition 6 (Nonstationary one-factor model)**.** *Let $\mathcal{G}$ be a DAG over latent variables $\mathbf{L} = \{\mathsf{L}_i\}^m$ where each latent variable L_i is generated following $\mathsf{L}_i = \sum_{\mathsf{L}_j \in Pa_{\mathcal{G}}(\mathsf{L}_i)} h_{j,i}(\mathsf{T}, \delta_{j,i})\mathsf{L}_j + g_i(\mathsf{T}, \epsilon_i)$. Each latent variable L_i is then associated with a set $\mathbf{X}_i$ of at least two observed variables (i.e., $|\mathbf{X}_i| \geq 2$) as its pure measurements. For each pure measurement $\mathsf{X}_i^{(k)} \in \mathbf{X}_i$, it is generated by $\mathsf{X}_i^{(k)} = h'_{i,k}(\mathsf{T}, \delta'_{i,k})\mathsf{L}_i + g'_{i,k}(\mathsf{T}, \epsilon'_{i,k})$, where we use primed notation (e.g., h') to distinguish parameters in the measurement process from those governing the latent variable dynamics.*

Note that in Phase 1, although the latent variables themselves are unobserved, the CI relations among them manifest as testable low-rank in observed measurements, due to the equivalent linearity in each domain (Theorem 4). In Phase 2, however, this rank-based correspondence breaks down, as changes to latent variables–represented by edges $\mathsf{T} \to \mathsf{L}_i$–can induce complex, nonlinear dependencies. Hence, instead of testing for CI relations between T and $\mathbf{L}$, we take a different route: parameter identification. We show that model parameters can be identified up to trivial indeterminacies in each domain, allowing us to detect changes by comparing identified parameters. We formalize this below.

We first introduce the model identification result. In each domain, there are following unknown model parameters: $\Sigma_{\mathbf{L},\mathbf{L}} \in \mathbb{R}^{m\times m}$, the variance-covariance matrix among $\mathbf{L}$; $\omega^{(1)}, \omega^{(2)}, \phi^{(1)}, \phi^{(2)} \in \mathbb{R}^m$, the equivalent linear coefficients and exogenous noise variances of each latent variable's two pure measurements, where for $k = 1, 2$, $\omega_i^{(k)} = \mathbb{E}[h'_{i,k}(t, \delta'_{i,k})]$, and $\phi_i^{(k)} = \mathrm{Var}[g'_{i,k}(t, \epsilon'_{i,k})]$. Let $\hat{\Sigma}_{\mathbf{L},\mathbf{L}}, \hat{\omega}^{(1)}, \hat{\omega}^{(2)}, \hat{\phi}^{(1)}, \hat{\phi}^{(2)}$ be another set of (estimated) parameters that yield the same covariance for $\mathbf{X}$. Then, the true parameters are identified up to scaling, i.e., $\frac{\hat{\omega}^{(1)}}{\omega^{(1)}} = \frac{\hat{\omega}^{(2)}}{\omega^{(2)}} =: c \in \mathbb{R}^m$, where the division is element-wise, and $\Sigma_{\mathbf{L},\mathbf{L}} = \mathrm{diag}(c)\, \hat{\Sigma}_{\mathbf{L},\mathbf{L}}\, \mathrm{diag}(c)$ holds. Moreover, measuring noise variances are identified, i.e., $\phi^{(k)} = \hat{\phi}^{(k)},\ k = 1, 2$. Proofs and solution details are given in Section B.

With the parameter identification results in each domain, we proceed to identify changes by comparing the recovered parameters across domains. Since latent variables are not accessible, this comparison must come with a trade-off. Two levels of assumptions are needed–one to address scaling indeterminacies, and one to ensure faithfulness so that changes leave trace in the second-order information:

Assumption 1. *Two assumptions are needed to identify changes from recovered model parameters:*

(A1.) To address indeterminacies, for each latent variable L_i, at least one of its measurement's equivalent linear coefficient, w.l.o.g. assumed to be $\omega_i^{(1)}$, remains invariant across domains.

(A2.) To ensure faithfulness, if the generating process of a latent variable L_i changes, then there exist at least two domains in which, for any subset $\mathbf{L}_C \subseteq \mathbf{L} \setminus \{\mathsf{L}_i\}$, the conditional variance of L_i given $\mathbf{L}_C$ (calculated from $\Sigma_{\mathbf{L},\mathbf{L}}$) differs. Similarly, if a measurement variable $\mathsf{X}_i^{(k)}$ is changed, its corresponding exogenous noise variance $\phi_i^{(k)}$ must change as well.

Due to space limit, we leave the explanation and justification of assumptions to Section B. Under the assumptions, we can now formally state the result for identifying changes, as in Theorem 5, to get the direct edges from T to $\mathbf{L}$ and $\mathbf{X}$. After that, further orientation can be done by using e.g., Meek rules.

Theorem 5 (Identification of changing variables)**.** *Suppose measurements $\mathbf{X}$ are generated following Definition 6 and let $\{\hat{\Sigma}_{\mathbf{L},\mathbf{L}|t}, \hat{\omega}_t^{(1)}, \hat{\omega}_t^{(2)}, \hat{\phi}_t^{(1)}, \hat{\phi}_t^{(2)}\}$ be the model parameters estimated in each domain $\mathsf{T} = t$. Denote the normalized latent covariance matrices by $\{\hat{\Sigma}'_{\mathbf{L},\mathbf{L}|t} \coloneqq \mathrm{diag}(\hat{\omega}_t^{(1)})\, \hat{\Sigma}_{\mathbf{L},\mathbf{L}|t}\, \mathrm{diag}(\hat{\omega}_t^{(1)})\}$. Then, under (A1) and (A2), $\mathsf{T} \to \mathsf{L}_i \in \mathcal{G}^{aug}$ if and only*

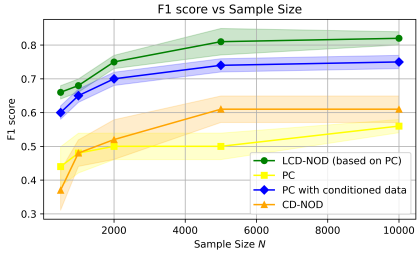


Figure 2: F1 score under the assumption of no latent confounders with 95% CI.

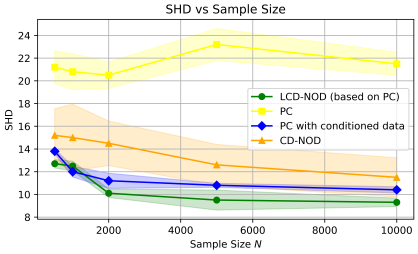


Figure 3: SHD under the assumption of no latent confounders with 95% CI.

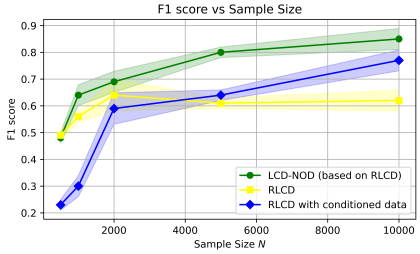


Figure 4: F1 under graphical assumptions required by RLCD with 95% CI.

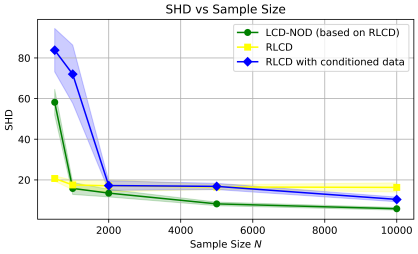


Figure 5: SHD under graphical assumptions required by RLCD with 95% CI.

if for all subsets $\mathbf{L}_C \subseteq \mathbf{L} \setminus \{\mathsf{L}_i\}$*, the conditional variances* $\{\hat{\mathrm{Var}}'_t(\mathsf{L}_i|\mathbf{L}_C)\}$ *calculated from* $\{\hat{\Sigma}'_{\mathbf{L},\mathbf{L}|t}\}$ *changes across* t*. And an* $\mathsf{T} \to \mathsf{X}_i^{(k)} \in \mathcal{G}^{aug}$ *if and only if the estimated* $\{\hat{\phi}_{i|t}^{(k)}\}$ *changes across* t*.*

5 Experiments

5.1 Synthetic Setting, Baselines, and Evaluation Metric

We employ synthetic data to validate the proposed rank test and LCD-NOD. Specifically, to simulate i.i.d. data from nonstationary POLCMs, we randomly generate DAG structures where each V_i has a 0.5 possibility to be influenced by nonstationarity, i.e., $h_{:,i}$ and g_i are a function of domain index $\mathsf{T} \in \{1, ..., 10\}$. For those that are not influenced by nonstationarity, the corresponding $h_{:,i}$ and g_i does not change across domain. The independent noise terms δ and ϵ are sampled from Gaussian with variance sampled uniformly from $[0.01, 0.1]$ and $[0.1, 1]$ respectively. For h and g, they are set as randomly initialized polynomial function of T and δ parameterized by neural network.

To validate the proposed conditional rank test, we employ the classical CCA-based rank test (Anderson, 1984) on the unconditional covariance matrix, as a baseline, referred to as Standard Rank Test, and also employ the standard rank test with conditional covariance matrix, referred to as Naive Conditional Rank Test. We refer to our method as Proposed Conditional Rank Test. We compare the Type-I and Type-II errors of each method. We consider six different sample sizes: 300, 500, 800, 1000, 1500, 2000, and adopt 10 random seeds to report the average performance.

To validate the proposed LCD-NOD, we employ three classical equality-constraint-based causal discovery methods, PCSpirtes et al. (2000), FOFC (Kummerfeld & Ramsey, 2016), and RLCD (Dong et al., 2024), and see whether the first phase of LCD-NOD can successfully upgrade these methods to handle the nonstationary scenario, in terms of F1 score (bigger better) and SHD (smaller better). These three methods consider structures without latent variables (Spirtes et al., 2000), structures of one factor model (Silva et al., 2003), and structures involving latent variables where all variables can be flexibly related (Dong et al., 2024). We also compare with CD-NOD (Huang et al., 2020b), which can also handle nonstationary data but cannot handle the existence of latent variables.

The Phase 2 of LCD-NOD is to determine which variables are directly influenced by T. Given that we allow the presence of latent variables, consider the whole structure, and allow T to influence both latent and observed variables, to our best knowledge, no existing method can achieve this result. Below is a comparison with the very related settings: CD-NOD must assume no latent variable, Jaber et al. (2020) only cares structure among observed and only allows interventions on observed, and LIT (Yang et al., 2024) only allow changes on exogenous noises and can only find edges from T to X. Thus, we mainly focus on comparing with the output of Phase 2 to ground truth. We focus on the end-2-end setting that take the data to LCD-NOD and get result of Phase 2 directly, as in Section 5.3, while also consider the disentangled setting detailed in Section A.3. Additionally, we compare the

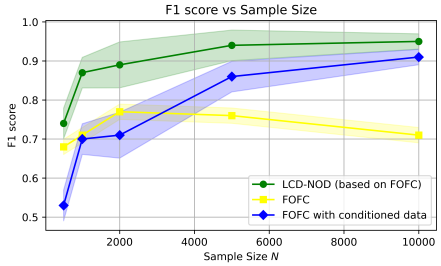


Figure 6: F1 score regarding $\mathcal{G}$ of each method under the OFM assumption with 95% CI.

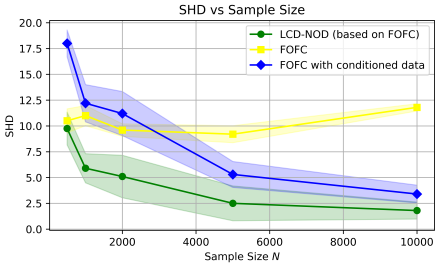


Figure 7: SHD regarding $\mathcal{G}$ of each method under the OFM assumption with 95% CI.

end-to-end performance of LCD-NOD with LIT (Yang et al., 2024) and UT-IGSP (Squires et al., 2020). To accommodate the constraints of LIT and UT-IGSP, this comparison is conducted under specific conditions that they require: allowing changes only in exogenous noises and identifying edges only from T to X. The result can be found in Section A.4, where LCD-NOD still consistently outperforms them though the setting is in favor of them.

5.2 Type-I and Type-II Comparison for the Proposed Conditional Rank Test

The result can be found in Figure 10 and Figure 11. It can be summarized that, both proposed conditional rank test and naive conditional rank test can properly control the Type-I errors (with significance level $\alpha = 0.05$). At the same time, the Type-II errors of our proposed method is consistently smaller than the baselines, which illustrates the better test power of our proposed method. More detailed analysis can be found in Section A.1.

5.3 LCD-NOD Performance on Synthetic Data

In this section, we employ synthetic data to validate the effectiveness of LCD-NOD. For Phase 1 of LCD-NOD, we compare PC, FOFC, and RLCD with their versions upgraded with LCD-NOD. We also compare with the naive conditioning upgrade strategy, referred to as PC with conditioned data, FOFC with conditioned data, and RLCD with conditioned data, respectively. We also compare with CD-NOD, which also handles nonstationarity but requires the absence of latent variables.

Figures 2 and 3 give the performance of each method in cases without latent variables. As shown, LCD-NOD consistently surpasses baselines in terms of both F1 score and SHD, across different sample sizes. The performance under the one factor model assumption (Silva et al., 2003), and under the graphical assumption required by RLCD (Dong et al., 2024) are given in Figures 6 and 7 and Figures 4 and 5, respectively. Similarly, LCD-NOD also outperforms all baselines, and the performance of LCD-NOD becomes better with the increase of sample size, while the original verson of PC, FOFC, and RLCD do not. This validates LCD-NOD as a general augmentation scheme of existing constraint-based causal discovery methods to handle the nonstationary scenario.

Last, we examine the end-2-end performance of Phase 2 of LCD-NOD. Specifically, we compare the output of Phase 2, i.e., the structure among $\mathbf{X}$, $\mathbf{L}$ and T, with the ground truth.The results are in Figures 12 and 13, where the Phase 2 performs quite well in terms of recovering the augmented structure, and benefits from the increase of sample size. E.g., when sample size is 2000 per domain, Phase 2 achieves 0.83 F1 and 6.5 SHD and it improves to 0.85 F1 and 5.7 SHD when the sample size becomes 5000. These results empirically validate the effectiveness of LCD-NOD. We also found that in Phase 2 it is easier to detect changes in observed variables. An intuitive explanation is as follows. A change in an observed variable, typically amounts to only a change in its measurement process, characterized by one set of conditional coefficients to estimate. However, a change in a latent variable can involve multiple other latent variables in the underlying structure, and hence multiple sets of coefficients conditioning on different latent variables need to be estimated. This is also aligned with the reason why we need to explicitly impose Assumption 1 A2 and to address the indeterminacies of latent variables to make such detection feasible.

5.4 LCD-NOD Performance on Real-world Data

We employ Big Five personality dataset (`openpsychometrics.org`) to show the real-life applicability of LCD-NOD. By taking country as the domain index, we found that causal mechanisms among certain dimensions, e.g., agreeableness and extroversion, indeed exhibits nonstationarity, which aligns with psychometric studies. Please kindly refer to Figure 9 and Section A.2.

6 Conclusion

This work formulates a class of nonstationary models, shows the constraints, and develops a principled test together with latent variable causal discovery method under nonstationarity.

REFERENCES

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34:22822–22833, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, 2023.

Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

Carlos Améndola, Mathias Drton, Alexandros Grosdos, Roser Homs, and Elina Robeva. Third-order moment varieties of linear non-gaussian graphical models. *Information and Inference: A Journal of the IMA*, 12(3):iaad007, 2023.

Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning*, pp. 249–257, 2013.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis. 2nd ed*. John Wiley & Sons, 1984.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Conference on Causal Learning and Reasoning*, 2024.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, 2020.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

Gonzalo Camba-Méndez and George Kapetanios. Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling. *Econometric Reviews*, 28(6):581–611, 2009.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In *Proceedings 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pp. 294–321, 2012.

Haoyue Dai, Peter Spirtes, and Kun Zhang. Independence testing-based approach to causal discovery under measurement error and linear non-gaussian models. *Advances in Neural Information Processing Systems*, 35:27524–27536, 2022.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representation*, 2024.

Mathias Drton. Algebraic problems in structural equation modeling. In *Advanced Studies in Pure Mathematics*, pp. 35–86. Mathematical Society of Japan, 2018.

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

Minghao Fu, Biwei Huang, Zijian Li, Yujia Zheng, Ignavier Ng, Yingyao Hu, and Kun Zhang. Identification of nonparametric dynamic causal structure and latent process in climate system. *arXiv preprint arXiv:2501.12500*, 2025.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(79): 2409–2464, 2012.

B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. In *JMLR*, volume 21(89), 2020a.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020b.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023. ISSN 2666-3899.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

Jikai Jin and Vasilis Syrgkanis. Learning causal representations from general environments: Identifiability and intrinsic ambiguity. *arXiv preprint arXiv:2311.12267*, 2023.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems*, 2019.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Erich Leo Lehmann, Joseph P Romano, et al. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Advances in Neural Information Processing Systems*, 2016.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.

Ignavier Ng, Xinshuai Dong, Haoyue Dai, Biwei Huang, Peter Spirtes, and Kun Zhang. Score-based causal discovery of latent variable causal models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ZdSe1qnuia.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

Thomas S Richardson. *Models of feedback: interpretation and discovery*. PhD thesis, Carnegie-Mellon University, 1996.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter L Spirtes. Learning measurement models for unobserved variables. *arXiv preprint arXiv:1212.2516*, 2003.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7 (2), 2006.

Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pp. 278–285. PMLR, 2001.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, 2023.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. *arXiv:0812.1938.*, 2010.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *arXiv preprint arXiv:2402.00849*, 2024a.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *arXiv preprint arXiv:2406.05937*, 2024b.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Conference on Uncertainty in Artificial Intelligence*, 1991.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, 2017.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

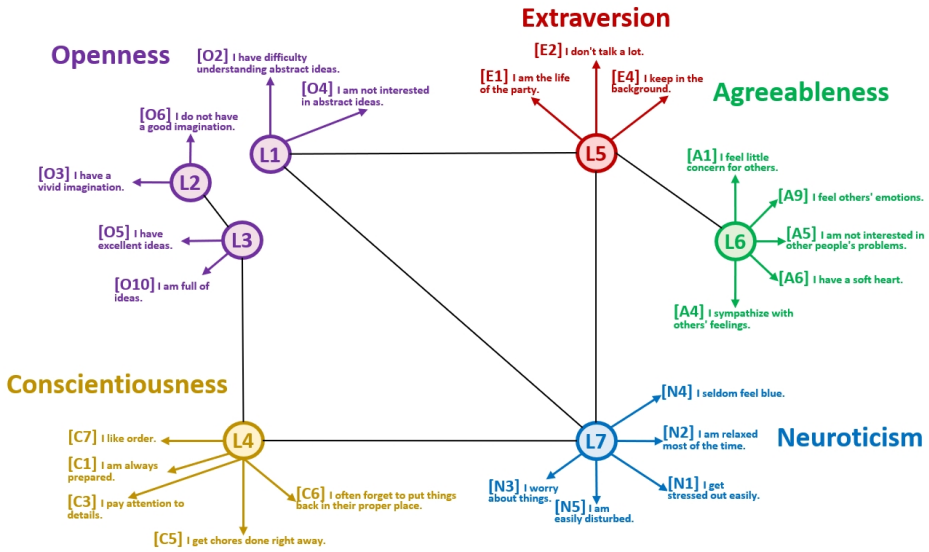
Yuqin Yang, Saber Salehkaleybar, and Negar Kiyavash. Learning unknown intervention targets in structural causal models from heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3187–3195. PMLR, 2024.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 2023.

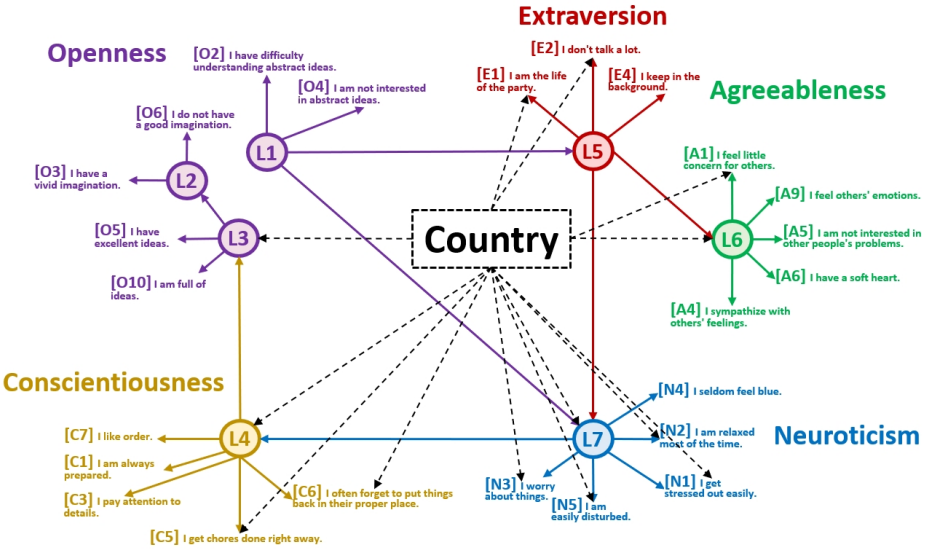
Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning*, 2024.

APPENDIX OF "IDENTIFYING PARTIALLY OBSERVED CAUSAL MODELS FROM HETEROGENEOUS/NONSTATIONARY DATA"



(a) Causal structure (CPDAG) on Big Five personality data found by Phase 1.



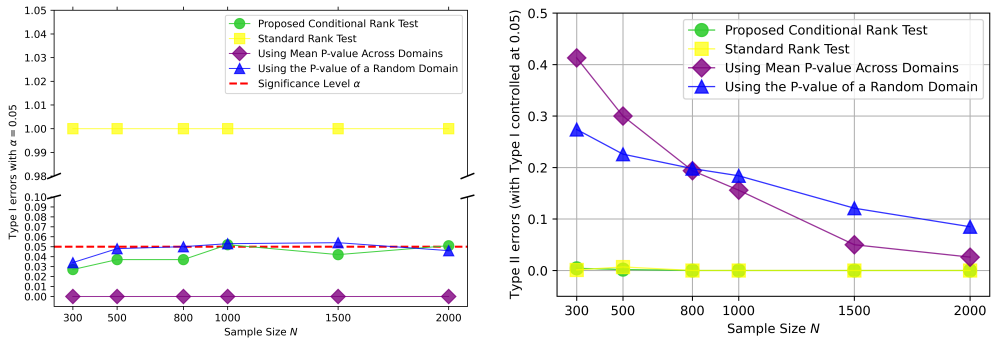
(b) The augmented structure (structure with domain index Country) on Big Five by Phase 2.

Figure 9: The causal structure found by LCD-NOD on Big Five personality data, where (a) is found by the Phase 1 of LCD-NOD and (b) is found by the Phase 2.

A ADDITIONAL EXPERIMENTAL RESULTS

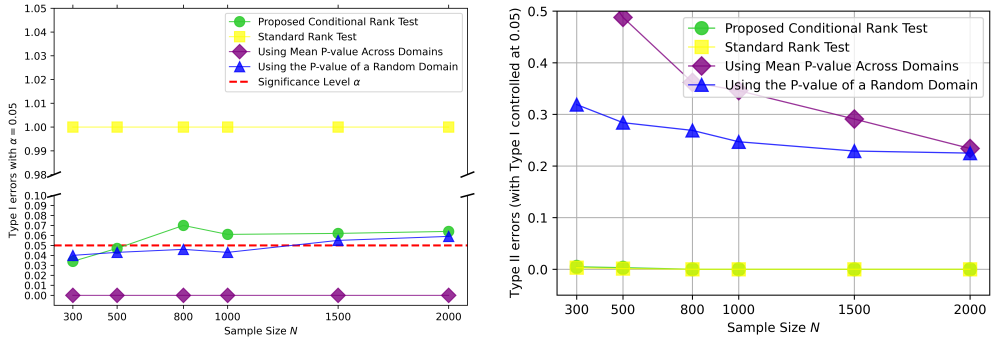
A.1 EXPERIMENTAL RESULTS FOR THE PROPOSED CONDITIONAL RANK TEST

In this section we empirically validate our proposed conditional rank test from two perspectives. 1. whether the proposed test is a valid test, by checking whether it can control the Type-I error properly



(a) The probability of Type I errors with $\alpha = 0.05$. (b) Type II errors (effective Type I controlled at 0.05).

Figure 10: The probability of Type I and Type II errors.



(a) The probability of Type I errors with $\alpha = 0.05$. (b) Type II errors (effective Type I controlled at 0.05).

Figure 11: The probability of Type I and Type II errors without assuming jointly gaussian.

at a designated significance level. 2. the power of the test, by comparing the Type-II error with baselines.

As for Type-I error, the result is shown in Figure 10 (a) (where variables are jointly gaussian conditioned on T) and Figure 11 (a) (without assuming jointly gaussian). Specifically, with significance level 0.05, the proposed conditional rank test can consistently control the Type-I errors around 0.05 with varying sample sizes, which shows that the proposed test is valid. Plus, if we use the standard rank test (which tests the unconditional rank), the Type-I error will be very large and close to 1. This also validates the motivation of our proposed novel conditional rank test for causal discovery in the nonstationary setting as directly plugging in a standard rank test will overly reject null hypotheses.

As for Type-II error, the result is shown in Figure 10 (b) (where variables are jointly gaussian conditioned on T) and Figure 11 (b) (without assuming jointly gaussian). As illustrated in the figures, the proposed conditional rank test can control the Type-II error very well, even when the joint Gaussianity is violated (Figure 11 (b)). We also note that both standard rank test and the proposed conditional rank test have a nearly zero Type-II error, even with a pretty small sample size ($N = 300$). However, the standard rank test achieves such a good Type-II error result, at the cost of having high Type-I error. In contrast. the proposed conditional rank test not only properly controls the Type-I error but also controls the Type-II error very effectively.

A.2 Results on Real-World Big Five Personality Dataset

In this section, we aim to employ the real-world Big Five personality dataset (`openpsychometrics.org`) to validate the proposed LCD-NOD method. This dataset contains 50 personality indicators / questions with 19,719 datapoints. Each data point corresponds to a person that participates the questionnaire and each indicator's value is the response ("Disagree",

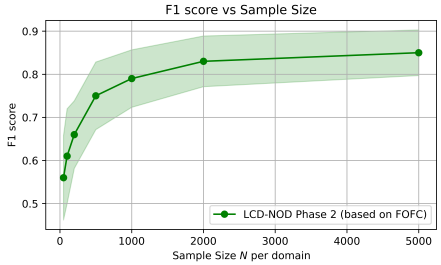


Figure 12: End-to-end performance of LCD-NOD Phase 2 by F1 score regarding $\mathcal{G}^{\text{aug}}$ (structure involving T) of each method under OFM.

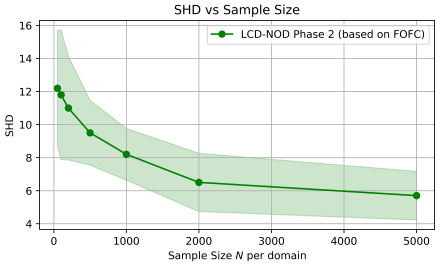


Figure 13: End-to-end performance of LCD-NOD Phase 2 by SHD regarding $\mathcal{G}^{\text{aug}}$ (structure involving T) of each method under OFM.

"Slightly Disagree", "Neutral", "Slightly Agree", "Agree") of the person to each question (e.g., "I am the life of the party"). Psychologists believe that there are five major dimensions that underlie human personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (O-C-E-A-N), and in the dataset, each dimension is designed to be measured by 10 questions (e.g., O1 is the first question for Openness). In this section we only use 24 out of the 50 questions to make the main conclusion of the result clearer. Further. the dataset contains the country information of the participants, which is taken as the domain index T in our experiment. As the number of data points for some countries could be very small, we choose the top-6 common countries in the dataset to produce the structure result, which are United States, United Kingdom, Canada, Australia, Philippines, and India. The structure produced by Phase 1 of LCD-NOD (based on FOFC) is shown in Figure 9 (a) and the augmented structure with country by Phase 2 of LCD-NOD is shown in Figure 9 (b).

As we can see, without any prior knowledge, the structure recovered by Phase 1 of our method aligns well with existing psychometric studies: each item in our result is indeed caused by the corresponding Big Five dimension, e.g., N1,...,N5 are all caused by the same latent variable L7, which is expected to correspond to Neuroticism. This result empirically validates the effectiveness of LCD-NOD from a psychological perspective.

We also found that, by using LCD-NOD with proposed conditional rank test, we can discovery meaningful structure with a reasonable significance level (1e-4); as a comparison, previous rank based methods such as hier-rank Huang et al. (2022) and RLCD Dong et al. (2024) have to use extremely small significance level (smaller than 1e-10) to produce plausible results. The reason lies in that, the underlying human personality model is heterogeneous and how variables affect each other varies across different countries. The previous rank based methods fail to consider and deal with such nonstationarity and their test p-values correspond to unconditional rank that cannot correctly reflect the desired t-separations. In contrast, by leveraging the proposed conditional rank test, LCD-NOD does not suffer from this issue.

Further, the Phase 2 of LCD-NOD can be leveraged to discovery how the underlying causal model changes across different countries. The corresponding result can be found in Figure 9 (b), where the edge from country to a variable, say, V_i, means that $h_{:,i}$ and g_i changes with country. Take L3 in Figure 9 (b) as an example. The existence of edge from country to L3 informs us that, although for all the countries, L4 (corresponds to conscientiousness) has a positive effect on L3 (a sub-dimension of openness that corresponds to ideas), the strength of such influence varies across countries. We further look into the edge coefficients and found that in India, the strength of the edge L4→L3 is around +0.45, which means a very strong causal influence from conscientiousness to ideativeness; on the contrary, in all other five countries this causal strength from L4 to L3 is only around 0.1. This informs us that, the underlying causal model related to variable L3 is indeed nonstationary across countries and especially different in India.

A.3 Disentangled Result for Phase 2

In this section we take the ground truth structure of the output of Phase 1 as the input of Phase 2. This is because, thought in the large sample limit, the phase 1 can produce the correct structure, given

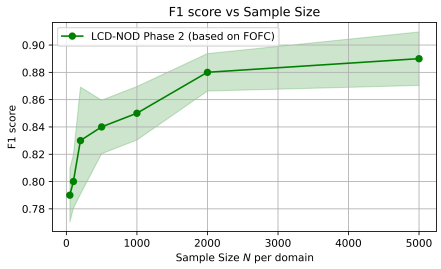


Figure 14: Disentangled performance of LCD-NOD Phase 2 F1 score regarding $\mathcal{G}^{\text{aug}}$ (structure involving T) of each method under the OFM assumption with 95% Confidence Interval.

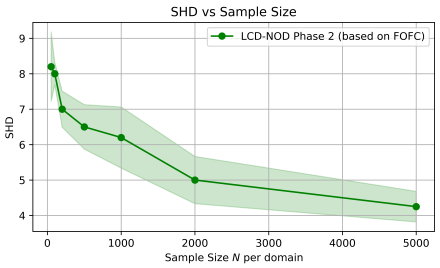


Figure 15: Disentangled performance of LCD-NOD Phase 2 by SHD regarding $\mathcal{G}^{\text{aug}}$ (structure involving T) of each method under the OFM assumption with 95% Confidence Interval.

Table 2: Comparison of the **end-to-end** performance of Phase 2 of LCD-NOD with LIT and UT-IGSP by F1 score with different sample sizes.

Method	Sample size per domain						
	N=50	N=100	N=200	N=500	N=1000	N=2000	N=10000
LCD-NOD (end-to-end)	**0.40**	**0.41**	**0.48**	**0.57**	**0.63**	**0.69**	**0.76**
LIT	0.27	0.24	0.31	0.28	0.31	0.32	0.32
UT-IGSP	0.30	0.36	0.42	0.39	0.40	0.39	0.41

finite samples, there always exists statistical errors. Thus this setting can test the performance of Phase 2 without the influence from the statistical errors in Phase 1. The result is shown in Figure 14 and Figure 15. As expected, LCD-NOD achieves better performance than that in the end-to-end setting.

A.4 End-to-end Comparison with LIT (Yang et al., 2024) and UT-IGSP (Squires et al., 2020)

Here we restrict the setting to accommodate the constraints of LIT and UT-IGSP. Specifically we allow changes only in exogenous noises and identifying edges only from T to X, and compare the **end-to-end** result of Phase 2 of LCD-NOD with LIT and UT-IGSP. The result is shown in Table 2. As shown in the table, LCD-NOD still consistently outperforms them across different sample sizes though the setting is in favor of them.

B Proofs

B.1 Proof of Theorem 1

Theorem 1 (Equivalence Relation between $\Theta(\mathcal{G})$ under POLCMs and $\Phi(\mathcal{G})$ under Nonstationary POLCMs)**.** *Given a DAG $\mathcal{G}$, let $\Theta(\mathcal{G})$ be the observational covariance set of $\mathcal{G}$ under POLCMS, as in Definition 3, and let $\Phi(\mathcal{G})$ be the observational conditional covariance set of $\mathcal{G}$ under Nonstationary POLCMs, as in Definition 4. We have $\Theta(\mathcal{G}) = \Phi(\mathcal{G})$.*

Proof. We first show the following Lemma 1.

Lemma 1 (Lemma for proof of Theorem 1)**.**

$$\mathbb{E}_{p(\mathbf{V}|\mathsf{T})}[(I - H(\mathsf{T})^T)^{-1} g(\mathsf{T}) g(\mathsf{T})^T (I - H(\mathsf{T})^T)^{-T}] \tag{9}$$

$$= \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[(I - H(\mathsf{T})^T)^{-1} g(\mathsf{T}) g(\mathsf{T})^T (I - H(\mathsf{T})^T)^{-T}] \tag{10}$$

$$= (I - \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})} H(\mathsf{T})^T)^{-1} \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[g(\mathsf{T}) g(\mathsf{T})^T] (I - \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})} H(\mathsf{T})^T)^{-T}. \tag{11}$$

Proof of Lemma 1. Let $\Omega_g = \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[gg^T]$. As g_i are mutually independent given T, Ω_g is a diagonal matrix. As g and H are independent given T, we have: $\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[(I - H^T)^{-1} g g^T (I -$

$H^T)^{-T}] = \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[(I-H^T)^{-1}\Omega_g(I-H^T)^{-T}]$. Now consider the i-th row and j-th column of it, i.e., $\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[(I-H^T)^{-1}\Omega_g(I-H^T)^{-T}]_{[i,j]} = \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[(I-H^T)^{-1}\Omega_g(I-H^T)^{-T}_{[i,j]}]$ $= \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[\sum_{P_1,P_2\in\mathcal{T}(\mathsf{V_i},\mathsf{V_j})}\Omega_{g\ \mathrm{top}(P_1,P_2)}H^{P_1}H^{P_2}]$, where $H^P = \Pi_{j\to i\in P}h_{j,i}(\mathsf{T},\delta_{j,i})$, $\mathcal{T}(\mathsf{V_i},\mathsf{V_j})$ refers to the set of treks between $\mathsf{V_i}$ and $\mathsf{V_j}$, and $\mathrm{top}(P_1,P_2)$ is the common source of the trek (P_1,P_2). As P_1,P_2 do not share edges, we have

$$\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[\sum_{P_1,P_2\in\mathcal{T}(\mathsf{V_i},\mathsf{V_j})}\Omega_{g\ \mathrm{top}(P_1,P_2)}H^{P_1}H^{P_2}] \tag{12}$$

$$= \sum_{P_1,P_2\in\mathcal{T}(\mathsf{V_i},\mathsf{V_j})}\Omega_{g\ \mathrm{top}(P_1,P_2)}\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H]^{P_1}\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H]^{P_2} \tag{13}$$

$$= (I-\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H]^T)^{-1}\Omega_g(I-\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H]^T)^{-T}. \tag{14}$$

□

Now, given Lemma 1, we prove Theorem 1.

(i) For every element of $\Theta(\mathcal{G})$ generated by (F,Ω), let $\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H] = F$ and $\mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[gg^T] = \Omega$. Then $\Phi(\mathcal{G})$ can generate the same element. (ii) For every element of $\Phi(\mathcal{G})$ generated by (H,g), let $F = \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[H]$ and $\Omega = \mathbb{E}_{p(\delta,\epsilon|\mathsf{T})}[gg^T]$. Then $\Theta(\mathcal{G})$ can generate the same element. Taking these two together, we have $\Theta(\mathcal{G}) = \Phi(\mathcal{G})$. □

B.2 Proof of Corollary 1

Corollary 1 (Equality and Inequality Constraints Under Nonstationary)**.** *If $\mathcal{G}$ implies an equality or inequality constraint on $\Theta(\mathcal{G})$, then $\mathcal{G}$ also implies the same constraint on $\Phi(\mathcal{G})$, and vice versa.*

Proof. If $\mathcal{G}$ implies a constraint on $\Theta(\mathcal{G})$, then by $\Theta(\mathcal{G}) = \Phi(\mathcal{G})$ as in Theorem 1, $\mathcal{G}$ also implies the same constraint on $\Phi(\mathcal{G})$, and vice versa. □

B.3 Proof of Theorem 2

Theorem 2 (Rank and T-separation for Nonstationary POLCMs)**.** *Given two sets of variables $\mathbf{A}$ and $\mathbf{B}$ generated by Nonstationary POLCMs with DAG $\mathcal{G}$ and assume rank faithfulness Dong et al. (2024). We have:*

$$\textit{rank}(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) = \min\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A},\mathbf{C_B})\ \textit{t-separates}\ \mathbf{A}\ \textit{from}\ \mathbf{B}\ \textit{in}\ \mathcal{G}\}, \tag{7}$$

where $\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}$ is the cross-covariance over $\mathbf{A}$ and $\mathbf{B}$ conditional on T.

Proof. By the relation between rank and t-separation in Sullivant et al. (2010) for stationary linear causal models, we have that $\mathrm{rank}(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) = \min\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A},\mathbf{C_B})$ t-separates $\mathbf{A}$ from $\mathbf{B}$ in $\mathcal{G}\}$. As rank constraints are a subset of equality constraints entailed by $\mathcal{G}$ for stationary linear causal models, we have $\mathrm{rank}(\Sigma_{\mathbf{A},\mathbf{B}})$ generated by Definition 2 equals $\mathrm{rank}(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}})$ generated by Definition 1. Thus, we have $\mathrm{rank}(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) = \min\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A},\mathbf{C_B})$ t-separates $\mathbf{A}$ from $\mathbf{B}$ in $\mathcal{G}\}$ for nonstationary POLCMs. □

B.4 Proof of Theorem 3

Theorem 3 (Likelihood Ratio Statistics for Rank of Conditional Covariance)**.** *Given two sets of variables $\mathbf{A}$ and $\mathbf{B}$ with $|\mathbf{A}| = P, |\mathbf{B}| = Q$ and $\mathbf{A}\cup\mathbf{B} = \mathbf{X}$ are jointly gaussian given T. Assume that the null hypothesis $\mathcal{H}_0$ is rank$(\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}) \le k$, the likelihood ratio statistics is as follows:*

$$\Lambda(k) = \sum_{t\in\textit{supp}(\mathsf{T})} -\left(N_t - \frac{P+Q+1}{2}\right)\ln(\Pi_{i=k+1}^{\min(P,Q)}(1-r_{t,i}^2)), \tag{8}$$

where N_t is the number of data points such that $\mathsf{T}=t$, and $r_{t,i}$ is the i-th canonical correlation between $\mathbf{A}$ and $\mathbf{B}$ conditioned on $\mathsf{T}=t$. We have that $\Lambda(k)$ converges in distribution to χ^2_{df}, with degree of freedom df $= \sum_{t\in\textit{supp}(\mathsf{T})}(P-k)(Q-k)$.

Proof. Let $D_{\mathbf{X}}$ be the observed data of $\mathbf{X}$ and $D^t_{\mathbf{X}}$ be the observed data of $\mathbf{X}$ conditioned on $\mathsf{T} = t$. The log-likelihood ratio statistics is:

$$\Lambda(k) = 2\log \frac{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} L(D_{\mathbf{X}}; \Sigma_{\mathbf{X}})}{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta} L(D_{\mathbf{X}}; \Sigma_{\mathbf{X}})} \tag{15}$$

$$= -2\log \frac{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} \Pi_{t \in \text{supp}(\mathsf{T})} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}}) P(\mathsf{T} = t)}{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta} \Pi_{t \in \text{supp}(\mathsf{T})} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}}) P(\mathsf{T} = t)} \tag{16}$$

$$= -2\log \frac{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} \Pi_{t \in \text{supp}(\mathsf{T})} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})}{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta} \Pi_{t \in \text{supp}(\mathsf{T})} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})} \tag{17}$$

$$= \sum_{t \in \text{supp}(T)} -2\log \frac{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})}{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})}. \tag{18}$$

For each t, the likelihood ratio statistics conditioned on t is $\lambda(k,t) = \frac{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} P(D^t_{\mathbf{X}}|\mathsf{T}=t;\Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})}{\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta} P(D^t_{\mathbf{X}}|\mathsf{T}=t;\Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})}$. The numerator $\sup_{\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}} \in \Theta_0} P(D^t_{\mathbf{X}}|\mathsf{T} = t; \Sigma_{\mathbf{X}|\mathsf{T}=\mathsf{t}})$ is a problem of MLE under rank constraint. By Anderson (1984); Muirhead (2009), the problem can be solved by calculating the empirical canonical correlation problem between the observations of $\mathbf{A}, \mathbf{B}$. More specifically, we have that $\lambda(k,t) = -\left(N_t - \frac{P+Q+1}{2}\right) \ln(\Pi_{i=k+1}^{\min(P,Q)}(1 - r^2_{t,i}))$, and thus Equation (8).

Now we show the asymptotic distribution of the statistics in Equation (8). As $\lambda(k,t)$ converges in distribution to $\chi^2_{(P-k)(Q-k)}$, and for different t, $\lambda(k,t)$ are mutually independent, by the use of continuous mapping theorem, $\Lambda(k) = \sum_{t \in \text{supp}(\mathsf{T})} \lambda(k,t)$ converges in distribution to $\chi^2_{\sum_{t \in \text{supp}(\mathsf{T})}(P-k)(Q-k)}$. □

B.5 Proof of Theorem 4

Theorem 4 (Structure Identifiability of LCD-NOD Phase 1)**.** *Given an equality-constraint-based causal discovery method $\mathcal{M}$, if $\mathcal{M}$ asymptotically identifies $\mathcal{G}$ up to equivalence class $\mathcal{C}$ under POLCMs and graphical assumption $\mathcal{A}$, then the Phase 1 of LCD-NOD, i.e., the augmented $\mathcal{M}$, asymptotically identifies $\mathcal{G}$ up to $\mathcal{C}$, under Nonstationary POLCMs and $\mathcal{A}$.*

Proof. In the large sample limit, the input to $\mathcal{M}$ and the augmented $\mathcal{M}$ are the population covariance over $\mathbf{X}$ and the conditional population covariance over $\mathbf{X}$ respectively. In other words, the inputs are an element of $\Theta(\mathcal{G})$ and an element of $\Phi(\mathcal{G})$ respectively. Under faithfulness, these two elements contain the same set of constraints as $\Theta(\mathcal{G})$ and $\Phi(\mathcal{G})$ respectively. By Corollary 1, these two elements contain the same set of equality constraints. Further, as $\mathcal{M}$ and the augmented $\mathcal{M}$ only makes use of equality constraints, the two algorithms will have exactly the same output, and thus the augmented $\mathcal{M}$ also asymptotically identifies $\mathcal{G}$ up to $\mathcal{C}$. □

B.6 Proof of Theorem 5

Theorem 5 (Identification of changing variables)**.** *Suppose measurements $\mathbf{X}$ are generated following Definition 6 and let $\{\hat{\Sigma}_{\mathbf{L},\mathbf{L}|t}, \hat{\omega}^{(1)}_t, \hat{\omega}^{(2)}_t, \hat{\phi}^{(1)}_t, \hat{\phi}^{(2)}_t\}$ be the model parameters estimated in each domain $\mathsf{T} = t$. Denote the normalized latent covariance matrices by $\{\hat{\Sigma}'_{\mathbf{L},\mathbf{L}|t} := \text{diag}(\hat{\omega}^{(1)}_t)\, \hat{\Sigma}_{\mathbf{L},\mathbf{L}|t}\, \text{diag}(\hat{\omega}^{(1)}_t)\}$. Then, under (A1) and (A2), $\mathsf{T} \to \mathsf{L}_i \in \mathcal{G}^{aug}$ if and only if for all subsets $\mathbf{L}_C \subseteq \mathbf{L} \setminus \{\mathsf{L}_i\}$, the conditional variances $\{\hat{\text{Var}}'_t(\mathsf{L}_i|\mathbf{L}_C)\}$ calculated from $\{\hat{\Sigma}'_{\mathbf{L},\mathbf{L}|t}\}$ changes across t. And an $\mathsf{T} \to \mathsf{X}^{(k)}_i \in \mathcal{G}^{aug}$ if and only if the estimated $\{\hat{\phi}^{(k)}_{i|t}\}$ changes across t.*

Proof. First, we show that using the indirect measurements from a single domain, the correspondences from latent variables to measurements can be determined, and the causal structure $\mathcal{G}$ among latent variables can be identified to its CPDAG. This is established in Corollary 1 and (Silva et al., 2003):

Suppose data is generated by the nonstationary one-factor model as in Definition 6, and assume rank faithfulness. The measurement clusters can first be identified, in that the cross-covariance matrix between two measured variables and all the remaining measured variables has a rank < 2 if and only if these two measured variables serve as measurements for a same latent variable. The CI relations among latent variables can then be identified with these clusters, in that any CI relation $\mathbf{L}_A \perp\!\!\!\perp \mathbf{L}_B | \mathbf{L}_C$ holds, if and only if the cross-covariance matrix between $\mathbf{X}_A \cup \mathbf{X}_C^{(1)}$ and $\mathbf{X}_B \cup \mathbf{X}_C^{(2)}$ is $|C|$, instead of higher. Here $\mathbf{X}_C^{(1)}$ and $\mathbf{X}_C^{(2)}$ are disjoint partitions of $\mathbf{X}_C$ such that $|\mathbf{X}_C^{(1)}|, |\mathbf{X}_C^{(2)}| \geq 2$. Then, with these CI relations recovered through ranks, one can apply e.g., PC as if having direct access to $\mathbf{L}$.

Then, we show that in additional to the model structure, it is also possible to consistently estimate (up to trivial indeterminacies) the model parameters from data in each domain. This allows us to estimate the conditional distributions $p(\mathsf{L}_i \mid \mathbf{L}_C, \mathsf{T} = t)$ directly, enabling the detection of changes:

Suppose we have access to measurements $\mathbf{X}$ in a single domain $\mathsf{T} = t$ generated by Definition 6. For simplicity, we drop the conditioning notations ($|\mathsf{T} = t$) when the scope is clear. There are following unknown model parameters: $\Sigma_{\mathbf{L},\mathbf{L}} \in \mathbb{R}^{m \times m}$ and $\mu_{\mathbf{L}} \in \mathbb{R}^m$, the variance-covariance matrix and the mean vector among $\mathbf{L}$; $\omega^{(1)}, \omega^{(2)}, \mu^{(1)}, \mu^{(2)}, \phi^{(1)}, \phi^{(2)} \in \mathbb{R}^m$, the equivalent linear coefficients, intercepts, and exogenous noise variances of each latent variable's two pure measuring processes, where for $k = 1, 2$, $\omega_i^{(k)} = \mathbb{E}[h'_{i,k}(t, \delta'_{i,k})]$, $\mu_i^{(k)} = \mathbb{E}[g'_{i,k}(t, \epsilon'_{i,k})]$, and $\phi_i^{(k)} = \mathrm{Var}[g'_{i,k}(t, \epsilon'_{i,k})]$.

To identify the model parameters, we first assume, without loss of generality, that each latent variable is dependent on at least one other latent variable. This assumption is necessary as otherwise its variance measurement parameters can be arbitrary. This assumption is also testable as isolated variables can be directly identified using marginal pairwise independence tests.

We further notice the trivial scaling and shifting indeterminacies: a latent variable can always be rescaled and shifted as long as its corresponding measurements are adjusted accordingly. However, we show that these are the only indeterminacies: Let $\hat{\Sigma}_{\mathbf{L},\mathbf{L}}, \hat{\mu}_{\mathbf{L}}, \hat{\omega}^{(1)}, \hat{\omega}^{(2)}, \hat{\mu}^{(1)}, \hat{\mu}^{(2)}, \hat{\phi}^{(1)}, \hat{\phi}^{(2)}$ be another set of model parameters (or estimators) that yield the same mean and covariance for $\mathbf{X}$. Let $\hat{\Sigma}_{\mathbf{L},\mathbf{L}}$ has unit diagonal, $\hat{\mu}_{\mathbf{L}} = 0$, and all entries of $\hat{\omega}^{(1)}$ are positive. Then, under these constraints, all remaining parameters are uniquely determined and can be expressed in closed form with the mean and covariance of $\mathbf{X}$. Specifically, they are:

- Off-diagonal entries of latent covariances: $\hat{\Sigma}_{\mathsf{L}_i,\mathsf{L}_j} = \mathrm{sign}(\Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_j^{(1)}}) \sqrt{\frac{\Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_j^{(1)}} \Sigma_{\mathsf{X}_i^{(2)},\mathsf{X}_j^{(2)}}}{\Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_i^{(2)}} \Sigma_{\mathsf{X}_j^{(1)},\mathsf{X}_j^{(2)}}}}$;
- Measuring weights: $\hat{\omega}_i^{(1)} = \sqrt{\frac{\Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_i^{(2)}} \Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_j^{(1)}}}{\Sigma_{\mathsf{X}_j^{(1)},\mathsf{X}_i^{(2)}}}}$, for any i, j with $\Sigma_{\mathsf{X}_j^{(1)},\mathsf{X}_i^{(2)}} \neq 0$;
- Measuring weights: $\hat{\omega}_i^{(2)} = \frac{\Sigma_{\mathsf{X}_i^{(1)},\mathsf{X}_i^{(2)}}}{\hat{\omega}_i^{(1)}}$;
- Measuring noise variances: $\hat{\phi}_i^{(k)} = \Sigma_{\mathsf{X}_i^{(k)},\mathsf{X}_i^{(k)}} - (\hat{\omega}_i^{(k)})^2$, for $k = 1, 2$, and
- Measuring noise means: $\hat{\mu}_i^{(k)} = \mathbb{E}(\mathsf{X}_i^{(k)})$, for $k = 1, 2$.

With these closed-form solutions, we have: the ratio between each latent variable's two measuring linear coefficients are identified, i.e., $\frac{\hat{\omega}^{(1)}}{\omega^{(1)}} = \frac{\hat{\omega}^{(2)}}{\omega^{(2)}} =: c \in \mathbb{R}^m$, where the division is element-wise. The latent covariance matrix is also identified to this scaling, satisfying $\Sigma_{\mathbf{L},\mathbf{L}} = \mathrm{diag}(c)\, \hat{\Sigma}_{\mathbf{L},\mathbf{L}}\, \mathrm{diag}(c)$. The means are identified to the shifting, i.e., $\omega^{(k)} * \mu_{\mathbf{L}} + \mu^{(k)} = \hat{\mu}^{(k)}$ holds for $k = 1, 2$. And last, measuring noise variances' exact values are identified, i.e., $\phi^{(k)} = \hat{\phi}^{(k)}$ holds for $k = 1, 2$.

Finally, with the identified model parameters (up to their indeterminacies) from each domain, we can compare them and identify the changing variables. Under faithfulness assumption, a latent variable L_i is changing, if and only if for all subsets $\mathbf{L}_C \subseteq \mathbf{L} \setminus \{\mathsf{L}_i\}$, the conditional distribution $\{p(\mathsf{L}_i | \mathbf{L}_C, \mathsf{T} = t)\}$ is changing with T. Using second-order information, this conditional distribution

can be characterized by the conditional variance $\mathrm{var}(\mathsf{L}_i|\mathbf{L}_C)$, and regression coefficients and intercept of L_i on $\mathbf{L}_C$. Assumption *(A2)* ensures that these second-order information is changed.

Then, from the estimated rescaled and shifted model parameters, with assumption *(A1)*, each latent variable L_i must have at least one measurement with invariant linear coefficients $\omega_i^{(k)}$ and means $\mu_i^{(k)}$ across domains. For each L_i, its unknown corresponding invariant measurement can be identified as follows: for each L_i, pick one measurement $\mathsf{X}_i^{(k)}$ as if it was invariant, and rescale and shift L_i from different domains so that $\hat{\omega}_i^{(k)}$ and $\hat{\mu}_i^{(k)}$ are the same across all domains. Then, use the parameters calibrated on this set of invariant measurements to determine and to count the number of "changes". Due to the minimal change principle, which is another way to put faithfulness, the choice of $\mathsf{X}_i^{(k)}$ that can achieve the minimum number of changes must correspond to the true invariant measurements, and those recovered changes must be the true changes. This is because, when parameters are calibrated on incorrectly specified, actually changing measurements, some other truly invariant parameters will be scaled/shifted to be changing, while for the true changes, they cannot be offset to invariances. □

C Additional Information

C.1 Definition of Trek

Definition 7 (Treks (Sullivant et al., 2010)). *In $\mathcal{G}$, a trek from X to Y is an ordered pair of directed paths (P_1, P_2) where P_1 has a sink X, P_2 has a sink Y, and both P_1 and P_2 have the same source Z.*

C.2 Implementation Details of LCD-NOD

Phase 1 of LCD-NOD is a general augmentation of existing equality-constraint-based methods, and its implementation details involving the details for the proposed conditional rank test and the details for how to substitute the test in the original equality-constraint-based methods with the proposed test. As for the proposed conditional rank test, each time we first have a null hypo that the rank of $\Sigma_{\mathbf{A},\mathbf{B}|\mathsf{T}}$ is smaller or equal to k, and follow Theorem 3 to calculate the test statistics $\Lambda(k)$ from observed data. Then we plug in the test statistics to the null distribution following Theorem 3 to calculate the p-value. Under a specific significance level α, we reject the null hypothesis when the p-value is smaller than α. In our synthetic data, we use $\alpha = 0.05$ for all the compared methods. As for how to substitute the original equality constraint test by the proposed conditional rank test, please kindly refer to Section 4.1.

For Phase 2, since we already have $\mathbf{L}$ to $\mathbf{X}$'s correspondence, theoretically we can directly solve for the parameters using the closed-form expressions provided in Section B.6. However, in practice, there might be model mis-specification and the input to the square root terms may not always be positive, i.e., some inequality constraints are not satisfied. Hence, we use MLE to estimate the parameters that produce the maximum likelihood for the observed sample covariances, though not necessarily the exact covariance values. Specifically, for each DAG over $\mathbf{L}$ consistent with the CPDAG obtained from Phase 1, we have one DAG over $\mathbf{L} \cup \mathbf{X}$. Using this DAG and observed data over $\mathbf{X}$, we identify the model parameters using the technique from (Dong parameter et al. 2024). Then, for each possible choice of invariant measurements, the corresponding parameters are calibrated and the changes are determined and counted. Finally, the configuration of DAGs and the set of invariant measurements that realize the minimum number of changes are identified as the equivalence of true DAGs under these changes, the true invariant measurements, and the corresponding changes are the true changes.

C.3 Runtime Analysis of LCD-NOD

First we note that in LCD-NOD, the runtime is almost irrelevant to the sample size, as we only need to calculate the covariance and conditional covariance matrices once and save it for further use; the result of the procedure is irrelevant to sample size in terms of time complexity.

For Phase 1, the time complexity depends on two things: the complexity of the proposed conditional rank test and the complexity of the to-be-upgraded baseline method. As for the rank test, the time complexity of the standard rank test for $\Sigma_{\mathbf{X},\mathbf{Y}}$ is $\mathcal{O}(\max(|\mathbf{X}|, |\mathbf{Y}|)^3)$ and the complexity of the proposed conditional rank test is $\mathcal{O}(|\mathrm{supp}(\mathsf{T})| \times \max(|\mathbf{X}|, |\mathbf{Y}|)^3)$, where $|\mathrm{supp}(\mathsf{T})|$ refers to the

number of different values that T can take. This is because we need to conditioned on each value of T to calculate the likelihood ratio statistics as in the proposed Theorem 3. We note that there is a trade-off between time complexity and test power. As likelihood ratio test is often asymptotically optimal in terms of power, we cannot further optimize the time complexity without compromising the test power. As a comparison, if we randomly choose a domain (a value of T), then we can also build a naive but valid conditional rank test that has the same complexity as that of the standard rank test; Yet, the power is not as good as the proposed likelihood ratio based conditional rank test (as shown in Figure 10 (b) and Figure 11 (b)). The time complexity for the to-be-upgraded equality-constraint-based methods, e.g., PC and RLCD, varies and highly depends on the number of variables and the sparsity of the ground truth graph. Thus, Specifically, in the worst case they have a complexity exponential in the number of variables. However, if the underlying graph is sparse, which is a common and reasonable assumption (Kalisch et al. 2007), the complexity becomes polynomial. In our empirical experiments with single Intel(R) Xeon(R) CPU E5-2470, the Phase 1 of LCD-NOD is pretty fast and it takes only around 10 seconds, 3 seconds, and 30 seconds, per graph with average 15 nodes for LCD-NOD (PC-based), LCD-NOD (FOFC-based), and LCD-NOD (RLCD-based) respectively.

For Phase 2, the time complexity depends on the number of MLE parameter estimations needed, that is, the number of DAGs over $\mathbf{L}$ consistent with the CPDAG estimated from Phase 1. This traversing process can be done in $\mathcal{O}(|\mathbf{L}|^4)$ using clique-picking and memorization (Wienobst et al. 2023). On each of the DAG, time complexity for parameter estimation is $\mathcal{O}(t|\mathbf{V}|^3)$, where $|\mathbf{V}|$ is the number of variables in the graph and t is the number of iterations of gradient descent (Sivan et al. 1997). Then finally, to choose the invariant measurement configuration, since all parameters are already identified up to indeterminacies, this becomes a simple rescaling of matrices and can be done at once by broadcasting the rescaling operations under different choices all directly to one tensor operation. In practice, since MLE with latent variables is non-convex and there may be local solutions, for each DAG we run the MLE estimation procedure under different learning rates and multiple restarts and choose the one with best likelihood. Even under this, for the O-C-E-A-N real dataset with 7 latent variables and 25 measurements, the whole time for identifying changes is less than 30 seconds, under the same experimental environment as mentioned above.

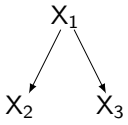
C.4 Illustrative Example to show The Relation between $\Theta(\mathcal{G})$ and $\Phi(\mathcal{G})$ Does not hold When g and h are not independent given T.

The example is shown in Figure 16. Specifically, (a) The causal graph of a model that satisfies Definition 1: $h_{j,i} = h_{j,i}(\mathsf{T}, \delta_{j,i})$ and $g_i = g_i(\mathsf{T}, \epsilon_i)$. Thus g and h are independent given T. (b) The causal graph that explicitly includes T, where $\mathsf{X}_2 \perp\!\!\!\perp \mathsf{X}_3 | \mathsf{T}, \mathsf{X}_1$ holds. (c) The same causal graph as (a) but the model does not satisfies Definition 1. Specifically, $h_{j,i} = h_{j,i}(\mathsf{T}, \delta_{j,i}, \mathsf{C})$ and $g_i = g_i(\mathsf{T}, \epsilon_i, \mathsf{C})$. Thus, g and h are not independent given T. (d) The effective causal graph that explicitly includes T and C, where X_1, T, and C are all the confounders of $\mathsf{X}_2, \mathsf{X}_3$, and thus the constraint $\mathsf{X}_2 \perp\!\!\!\perp \mathsf{X}_3 | \mathsf{T}, \mathsf{X}_1$ does not hold any more.

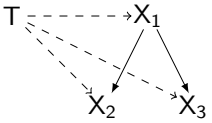
D Related Work

Causal discovery with latent variables: Early works that accommodate latent confounders were the FCI method (Spirtes et al., 2000; Richardson & Spirtes, 2002; Zhang, 2008), which utilize conditional independence tests to identify causal relations among observed variables while accounting for latent confounders. While FCI does not make any assumption about the latent structure, it often produces less informative outputs, e.g., provides no information about relationships among latent variables. On the other hand, FCI has shown to be maximally informative when considering nonparametric conditional independence constraints.

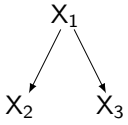
Therefore, to go beyond them, further constraints or information have been leveraged, such as those relying on parametric assumptions, as discussed in Section 1. A common approach is to use rank constraints (Sullivant et al., 2010), a generalization of the classical Tetrad constraints (Spirtes et al., 2000) and conditional independence constraints. This leads to a number of works that rely on different graphical assumptions (Silva et al., 2003; Huang et al., 2022; Dong et al., 2024). Alternatively, several other methods also rely on higher-order information (Shimizu et al., 2009; Cai et al., 2019; Salehkaleybar et al., 2020; Xie et al., 2020; Adams et al., 2021).



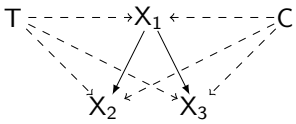
(a) The causal graph for a model that satisfies Definition 1.



(b) $\mathcal{G}^{\text{aug}}$ under nonstationary setting where the dashed arrows from T to variables means g and h are funcstion of T.



(c) The causal graph for a model where g and h are functions of both T and C.



(d) The corresponding causal graph that explicitly includes T and C, where X_1, T, and C are all the confounders of $\mathsf{X}_2, \mathsf{X}_3$, and thus the constraint that $\mathsf{X}_2 \perp\!\!\!\perp \mathsf{X}_3 | \mathsf{T}, \mathsf{X}_1$ does not hold any more.

Figure 16: Illustrative example to show that, if g and h are not independent given T, the constraints on the conditional covariance matrix (conditioned on T) will be different.

Causal discovery from heterogeneous data: Score-based methods have been developed to infer causal structure from heterogeneous data, which include those based on greedy search (Hauser & Bühlmann, 2012; Squires et al., 2020) or continuous optimization (Brouillard et al., 2020). On the other hand, Huang et al. (2020a) developed a constraint-based method that relies on conditional independence test, while Mooij et al. (2020) proposed a general framework that can incorporate different causal discovery methods.

Causal discovery with latent variables and causal representation learning from heterogenous data: Magliacane et al. (2016); Kocaoglu et al. (2019) proposed constraint-based methods that rely on conditional independence tests to recover ancestral structures over the observed variables, similar in spirit to FCI. In contrast, a related line of work, causal representation learning (Schölkopf et al., 2021), aims to infer both the latent causal variables and the causal structure among them. A special case of causal representation learning is nonlinear ICA which assumes that the latent variables are independent (Hyvarinen & Morioka, 2017; 2016; Hyvarinen et al., 2019; Hyvärinen et al., 2023). These methods also leverage interventional or heterogeneous data, such as single-node interventions (Ahuja et al., 2023; Squires et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Varıcı et al., 2024a) or multi-node interventions (Jin & Syrgkanis, 2023; Zhang et al., 2024; Varıcı et al., 2024b; Bing et al., 2024; Ng et al., 2025). Furthermore, some of them require hard interventions, such as von Kügelgen et al. (2023); Bing et al. (2024). Note that this line of approaches based on causal representation learning typically make certain assumptions: (1) no causal edges exist among observed variables or from observed to latent variables, and (2) the generative process from latent variables is deterministic (except several works including Khemakhem et al. (2020); Lachapelle et al. (2024); Fu et al. (2025)) and invariant across domains. In our work, we consider a more general setting that relaxes these assumption.

E LIMITATIONS

One limitation of this work is that, our proposed conditional rank test has to assume that all variables $\mathbf{X}$ are jointly gaussian given T; otherwise it is very difficult to derive the null distribution. However, we note that this is a common limitation of existing rank test, as the standard rank test also has to assume jointly gaussian. Plus, our empirical result in Section A.1 (Figure 11) empirically shows that, even when the data is not jointly gaussian, the proposed method can still control the Type-I properly and control the Type-II error effectively. A permutation-based rank test might surpass this jointly gaussian assumption and we plan to leave it for future exploration.

F BROADER IMPACTS

The goal of this paper is to advance the field of machine learning. We do not see any potential negative societal impacts of the work.