Journal of Machine Learning Research 21 (2020) 1-53 Submitted 3/19; Revised 2/20; Published 5/20

Causal Discovery from Heterogeneous/Nonstationary Data

Biwei Huang * BIWEIH@ANDREW.CMU.EDU
Kun Zhang * KUNZ1@CMU.EDU
Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213

Jiji Zhang JIJIZHANG@LN.EDU.HK
Department of Philosophy, Lingnan University, Tuen Mun, Hong Kong

Joseph Ramsey JDRAMSEY@ANDREW.CMU.EDU
Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213

Ruben Sanchez-Romero RUBEN.SARO@RUTGERS.EDU
Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ 07102

Clark Glymour CG09@ANDREW.CMU.EDU
Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany

Editor: Jon McAuliffe

arXiv:1903.01672v4 [cs.LG] 18 Jun 2020

Abstract

It is commonplace to encounter heterogeneous or nonstationary data, of which the underlying generating process changes across domains or over time. Such a distribution shift feature presents both challenges and opportunities for causal discovery. In this paper, we develop a framework for causal discovery from such data, called Constraint-based causal Discovery from heterogeneous/NOnstationary Data (CD-NOD), to find causal skeleton and directions and estimate the properties of mechanism changes. First, we propose an enhanced constraint-based procedure to detect variables whose local mechanisms change and recover the skeleton of the causal structure over observed variables. Second, we present a method to determine causal orientations by making use of independent changes in the data distribution implied by the underlying causal model, benefiting from information carried by changing distributions. After learning the causal structure, next, we investigate how to efficiently estimate the "driving force" of the nonstationarity of a causal mechanism. That is, we aim to extract from data a low-dimensional representation of changes. The proposed methods are nonparametric, with no hard restrictions on data distributions and causal mechanisms, and do not rely on window segmentation. Furthermore, we find that data heterogeneity benefits causal structure identification even with particular types of confounders. Finally, we show the connection between heterogeneity/nonstationarity and soft intervention in causal discovery. Experimental results on various synthetic and real-world data sets (task-fMRI and stock market data) are presented to demonstrate the efficacy of the proposed methods.

Keywords: causal discovery, heterogeneous/nonstationary data, independent-change principle, kernel distribution embedding, driving force estimation, confounder

∗. Equal contribution

©2020 Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v21/19-232.html.

1. Introduction

Many tasks across several disciplines of empirical sciences and engineering rely on the underlying causal information. As it is often difficult, if not impossible, to carry out randomized experiments, inferring causal relations from purely observational data, known as the task of causal discovery, has drawn much attention in machine learning, philosophy, statistics, and computer science. Traditionally, for causal discovery from observational data, under appropriate assumptions, so-called constraint-based approaches recover some information of the underlying causal structure based on conditional independence relationships of the variables (Spirtes et al., 1993). Alternatively, approaches based on functional causal models infer the causal structure by exploiting the fact that under certain conditions, the independence between the noise and the hypothetical cause only holds for the correct causal direction but not for the wrong direction (Shimizu et al., 2006; Hoyer et al., 2009; Zhang and Hyvärinen, 2009a,b).

Over the last few years, with the rapid accumulation of huge volumes of data of various types, causal discovery is facing exciting opportunities but also great challenges. One feature such data often exhibit is distribution shift. Distribution shift may occur across data sets, which may be obtained under different interventions or with different data collection conditions, or over time, as featured by nonstationary data. For an example of the former kind, consider remote sensing imagery data. The data collected in different areas and at different times usually have different distributions due to varying physical factors related to ground, vegetation, illumination conditions, etc. As an example of the latter kind, fMRI recordings are usually nonstationary: information flows in the brain may change with stimuli, tasks, attention of the subject, etc. More specifically, it is believed that one of the basic properties of neural connections is their time-dependence (Havlicek et al., 2011). In these situations many existing approaches to causal discovery may fail, as they assume a fixed causal model and hence a fixed joint distribution underlying the observed data. For example, if changes in local mechanisms of some variables are related, one can model the situation as if there exists some unobserved quantity which influences all those variables and, as a consequence, the conditional independence relationships in the distribution-shifted data will be different from those implied by the true causal structure.

In this paper, we assume that mechanisms or parameters, associated with the causal model, may change across data sets or over time (we allow mechanisms to change in such a way that some causal links in the structure may vanish or appear in some domains or over some time periods). We aim to develop a principled framework to model such situations as well as practical methods, called Constraint-based causal Discovery from heterogeneous/NOnstationary Data (CD-NOD), to address the following questions:

1. How can we efficiently identify variables with changing local mechanisms and reliably recover the skeleton of the causal structure over observed variables?
2. How can we take advantage of the information carried by distribution shifts for the purpose of identifying causal directions?

After identifying the causal structure, it is then appealing to ask how causal mechanisms change across domains or over time, which raises the question:

3. How can we extract from data a low-dimensional and potentially interpretable representation of changes, the so-called "driving force" of changing causal mechanisms?

Furthermore, we extend our approach to deal with more general scenarios, e.g., dynamic systems which involve both time-varying instantaneous and lagged causal relations and the case of stationary confounders.

In answering these questions, we make use of the following properties of causal systems. (i) Causal models and distribution shifts are heavily coupled: causal models provide a compact description of how data-generating processes, as well as data distributions, change, and distribution shifts exhibit such changes. (ii) From a causal perspective, the distribution shift in heterogeneous/nonstationary data is usually constrained—it may be due to the changes in the data-generating processes (i.e., the local causal mechanisms) of *a small number* of variables. (iii) From a latent variable modeling perspective, heterogeneous/nonstationary data are generated by some quantities that change across domains or over time, which gives hints as to how to understand distribution shift and estimate its underlying driving forces. (iv) Suppose that there are no confounders for `cause` and `effect`. Then $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ are either fixed or change independently, also known as the modularity property of causal systems (Pearl, 2000). Such an independence helps identify causal directions in the presence of distribution shifts.

To reliably estimate the skeleton of the causal structure and detect changing causal mechanisms from heterogeneous/nonstationary data (Problem 1), we make use of property (i), (ii), and (iii) listed above. Specifically, we introduce a surrogate variable C into the causal system to characterize hidden quantities that lead to the changes across domains or over time. The variable C can be a domain or time index. Including C in the causal system provides a convenient way to unpack distribution shifts to causal representations. We show that given C, (conditional) independence relationships between observed variables are the same as those implied by the true causal structure. We, additionally, show that variables that are adjacent to the surrogate variable C have changing causal mechanisms. We make the assumption of faithfulness on the graph involving C, and it is known that faithfulness implies a minimality condition on the edges (Zhang and Spirtes, 2016); as a consequence, the graphical representation produced by our procedure naturally *enjoys a "minimal change" principle*—the representation explains the conditional independence relations and changeability of the distribution with a minimal number of changing conditional distributions.

Regarding Problem 2, as a sub-problem of causal discovery, we show that distribution shift provides additional information for causal direction identification. it is known that with functional causal model-based approaches, there are cases where causal directions are not identifiable, e.g., the linear-Gaussian case and the case with a general functional class (Hyvärinen and Pajunen, 1999; Zhang et al., 2015b). This restricts the causal direction identification to certain functional classes, e.g., additive (Shimizu et al., 2006; Hoyer et al., 2009; Zhang and Hyvärinen, 2009a) or post-nonlinear models (Zhang and Chan, 2006; Zhang and Hyvärinen, 2009b). We show that using information carried by distribution shifts does not suffer from these restrictions—the method applies to general causal mechanisms.

Specifically, we take advantage of property (iv) for causal direction determination: if there is no confounder for V_i and V_j, then the causal mechanisms, represented by the conditional distributions $P(V_i \mid \mathrm{PA}^i)$ and $P(V_j \mid \mathrm{PA}^j)$, change independently across data sets or over time. However, independence typically no longer holds for wrong directions. This gives rise to causal asymmetry. To exploit this asymmetry, we develop a kernel embedding

of nonstationary conditional distributions to represent changing causal mechanisms and accordingly propose a dependence measure to determine causal directions. Furthermore, it is worth noting that although our method can been seen as an extension of constraint-based methods such as PC (Spirtes et al., 2001), unlike the original ones, it can *find the causal direction between even two variables*, thanks to the surrogate variable C in the system or more generally, the independent change property implied by a causal system.

Regarding Problem 3, traditionally, one may use Bayesian change point detection to detect change points of observed time series (Adams and MacKay, 2007) or use sliding windows and then estimate the causal model within each segment separately. However, Bayesian change point detection can only be applied to detect changes in marginal or joint distributions, whereas causal mechanisms are represented by conditional distributions. Moreover, neither of them is appropriate when causal mechanisms change continuously over time. More recently, a window-free method has been proposed, by extending Gaussian process regression (Huang et al., 2015). However, it requires the assumption of linearity, and it fails to handle the case when nonstationarity results from the change of noise distributions. In this paper, by leveraging property (iii), we propose a nonparametric method to *recover a low-dimensional and interpretable representation of mechanism changes*, which does not rely on window segmentation.

This paper is organized as follows.[1] In Section 2 we define and motivate the problem in more detail and review related work. Section 3 proposes an enhanced constraint-based method to recover the causal skeleton over observed variables and identify variables with changing causal mechanisms. Section 4 develops a method for determining causal directions by exploiting distribution shifts. It makes use of the property that in a causal system, causal modules change independently if there are no confounders. Section 5 proposes a method, termed Kernel Nonstationary Visualization (KNV), to visualize a low-dimensional and interpretable representation of changing mechanisms, the so-called "driving force". In Section 6, we extend CD-NOD to the case that allows both time-varying lagged and instantaneous causal relationships, and we discuss whether distribution shifts also help for causal discovery when there exist stationary confounders. In addition, we give a procedure to leverage both CD-NOD and approaches based on constrained functional causal models. In Section 7, we show the connection between heterogeneity/nonstationarity and soft intervention in causal

1. This paper is built on the arXiv paper by Zhang et al. (2015a) and the conference papers by Zhang et al. (2017) and Huang et al. (2017) but is significantly extended in several aspects. We reformulate assumptions in Section 3.1. We extend Section 3.2 to show how we detect pseudo confounders for nonadjacent variables. In Section 4.1, we add Algorithm 2 which uses generalization of invariance to identify causal directions. In Section 4.2, we propose a new approach to efficiently identify causal directions using independent changes between causal mechanisms and detect pseudo confounders behind adjacent variables (Algorithm 3). In Section 4.3, we give identifiability conditions of CD-NOD and define the equivalence class that CD-NOD can achieve if those conditions do not hold. Furthermore, we extend CD-NOD to the case when there exist both time-varying instantaneous and lagged causal relationships (Section 6.1). Accordingly, we propose Algorithm 5 to efficiently recover both instantaneous and time-lagged causal relationships. In Section 6.2, we further discuss whether distribution shifts also help for causal discovery when there exist stationary confounders. With CD-NOD, some causal directions may not be identifiable, if the identifiability conditions are not satisfied (Theorem 2). To make the method more applicable, we combine our framework with approaches based on constrained functional causal models (Section 6.3). In Section 7, we show that heterogeneity/nonstationarity and soft intervention are related in causal discovery, and we find that our proposed method is even more effective.

discovery. Section 8 reports experimental results tested on both synthetic and real-world data sets, including task-fMRI data, Hong Kong stock data, and US stock data.

2. Problem Definition and Related Work

In this section, we first review causal discovery approaches with fixed causal models. Then we give examples to show that if the underlying causal model changes, directly applying approaches with fixed causal models may result in spurious edges or wrong causal directions, which motivates our work in causal discovery with changing causal models.

2.1 Causal Discovery of Fixed Causal Models

Most causal discovery methods assume that there is a fixed causal model underlying the observed data and aim to estimate it from the data. Classic approaches to causal discovery divide roughly into two types. In the late 1980's and early 1990's, it was noted that under appropriate assumptions, one could recover a Markov equivalence class of the underlying causal structure based on conditional independence relationships among the variables (Spirtes et al., 1993). This gave rise to the constraint-based approach to causal discovery, and the resulting equivalence class may contain multiple DAGs (or other graphical objects to represent causal structures), which entail the same conditional independence relationships. The required assumptions include the causal Markov condition and the faithfulness assumption, which entail a correspondence between d-separation properties in the underlying causal structure and statistical independence properties in the data. The so-called score-based approach (see, e.g., Chickering, 2003; Heckerman et al., 1995) searches for the equivalence class which gives the highest score under some scoring criteria, such as the Bayesian Information Criterion (BIC), the posterior of the graph given the data (Heckerman et al., 2006), and the generalized score functions (Huang et al., 2018).

Another set of approaches is based on constrained functional causal models, which represent the effect as a function of the direct causes together with an independent noise term. The causal direction implied by the constrained functional causal model is generically identifiable, in that the model assumptions, such as the independence between the noise and cause, hold only for the true causal direction and are violated for the wrong direction. Examples of such constrained functional causal models include the linear non-Gaussian acyclic model (LiNGAM, Shimizu et al., 2006), the additive noise model (Hoyer et al., 2009; Zhang and Hyvärinen, 2009a), and the post-nonlinear causal model (Zhang and Chan, 2006; Zhang and Hyvärinen, 2009b).

2.2 With Changing Causal Models

Suppose that we are given a set of observed variables $\mathbf{V} = \{V_i\}_{i=1}^m$ whose causal structure is represented by a DAG G. For each V_i, let PA^i denote the set of parents of V_i in G. Suppose that at each time point or in each domain, the joint probability distribution of $\mathbf{V}$ factorizes according to G:

$$P(\mathbf{V}) = \prod_{i=1}^{m} P(V_i \,|\, \mathrm{PA}^i). \tag{1}$$



Figure 1: An illustration on how ignoring changes in the causal model may lead to spurious edges by constraint-based methods. (a) The true causal graph (including confounder $g(C)$, which is hidden). (b) The estimated causal skeleton on the observed data in the asymptotic case given by constraint-based methods, e.g., PC or FCI.

We call each $P(V_i \,|\, \mathrm{PA}^i)$ a *causal module* (the same meaning with "causal mechanism" in previous sections). If there are distribution shifts (i.e., $P(\mathbf{V})$ changes across domains or over time), at least some causal modules $P(V_k \,|\, \mathrm{PA}^k)$, $k \in \mathcal{N}$ must change. We call those causal modules *changing causal modules*. Their changes may be due to changes of the involved functional models, causal strengths, noise levels, etc. We assume that those quantities that change across domains or over time can be written as functions of a domain or time index, and denote by C such an index.

If the changes in some modules are related, one can treat the situation as if there exists some unobserved quantity (confounder) which influences those modules and, as a consequence, the conditional independence relationships in the distribution-shifted data will be different from those implied by the true causal structure. Therefore, standard constraint-based algorithms such as PC or FCI (Spirtes et al., 1993) may not be able to reveal the true causal structure. As an illustration, suppose that the observed data were generated according to Fig. 1(a), where $g(C)$, a function of C, is involved in the generating processes for both V_2 and V_4; the causal skeleton over the observed data then contains spurious edges $V_1 - V_4$ and $V_2 - V_4$, as shown in Fig. 1(b), because there is only one conditional independence relationship, $V_3 \perp\!\!\!\perp V_1 \,|\, V_2$.

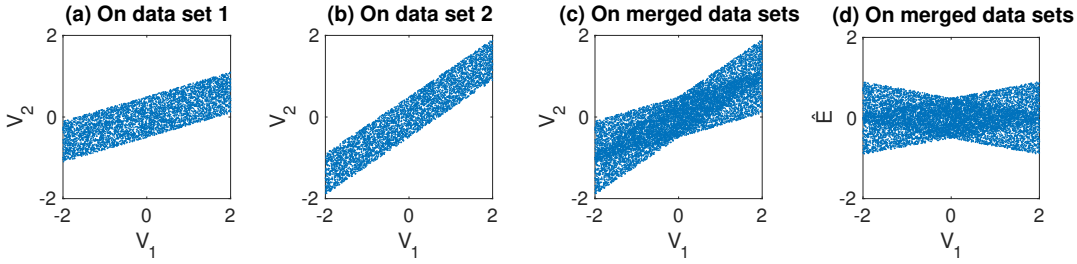


Figure 2: An illustration of a failure of using the approach based on functional causal models for causal direction determination when the causal model changes. (a) Scatter plot of V_1 and V_2 on data set 1. (b) Scatter plot of V_1 and V_2 on data set 2. (c) Scatter plot of V_1 and V_2 on merged data (both data sets). (d) Scatter plot of V_1 and the estimated regression residual $\hat{E}$ on merged data by regressing V_2 on V_1.

Moreover, when one fits a fixed functional causal model (e.g., a linear, non-Gaussian model, Shimizu et al., 2006) to distribution-shifted data, the estimated noise may not be independent of the cause. Consequently, the approach based on constrained functional causal models, in general, cannot infer the correct causal structure either. Figure 2 gives an illustration of this point. Suppose that we have two data sets for variables V_1 and V_2: V_2 is generated from V_1 according to $V_2 = 0.3V_1 + E$ in the first data set and according to $V_2 = 0.7V_1 + E$ in the second one, and in both data sets V_1 and E are mutually independent and follow a uniform distribution. Figure 2(a-c) show the scatter plots of V_1 and V_2 on data set 1, on data set 2, and on merged data, respectively. Figure 2(d) shows the scatter plot of V_1, the cause, and the estimated regression residual on the merged data set by regressing V_2 on V_1; they are not independent anymore, although on either data set the regression residual is independent of V_1. Thus, we cannot correctly determine the causal direction.

To tackle the issue of changing causal models, one may try to find causal models in each domain, for data from multiple domains, or in each sliding window, for nonstationary data, separately, and then compare and merge them. For instance, regarding the former type of data from multiple domains, in particular, multiple data sets obtained by external interventions where it is unknown what variables are manipulated, He and Geng (2016) considered two settings, depending on the sample size of each data set–in the setting with a large sample size for each data set, they proposed a graph-merging method after learning a causal network in each domain separately; in the setting with a relative small sample size, they proposed to pool together the data to learn a network structure and then use a re-sampling approach to evaluate the edges of the learned network. Regarding nonstationary data, improved versions include the online change point detection method (Adams and MacKay, 2007), the online undirected graph learning (Talih and Hengartner, 2005), and the locally stationary structure tracker algorithm (Kummerfeld and Danks, 2013). Such methods may suffer from high estimation variance due to sample scarcity, large type II errors, or multiple testing problems from a large number of statistical tests. Some methods aim to estimate the time-varying causal model by making use of certain types of smoothness of the change (Huang et al., 2015), but they do not explicitly locate the nonstationary causal modules. Several methods aim to model time-varying time-delayed causal relations (Xing et al., 2010; Song et al., 2009), which can be reduced to online parameter learning because the direction of causal relations is given (i.e., the past influences the future). Compared to them, learning changing instantaneous causal relations, with which we are concerned in this paper, is generally more difficult. Recently, several methods have been proposed to tackle time-varying or domain-varying instantaneous causal relations (Ghassami et al., 2018; Huang et al., 2019a,b, 2020). However, they assume linear causal models, limiting their applicability to complex problems with nonlinear causal relations.

In contrast, we develop a nonparametric and computationally efficient method that can identify changing causal modules and reliably recover the causal structure. We show that distribution shifts actually contain useful information for the purpose of determining causal directions and develop practical algorithms accordingly. After identifying the causal structure, we propose a method to estimate a low-dimensional and interpretable representation of changes.

3. CD-NOD Phase I: Changing Causal Module Detection and Causal Skeleton Estimation

In this section, we first formalize the assumptions that will be used in CD-NOD. Specifically, we allow a particular type of confounders—pseudo confounders, and we do not put hard restrictions on functional forms of causal mechanisms and data distributions. Accordingly, we propose an approach to efficiently detect changing causal modules and identify the causal skeleton; we call this step CD-NOD phase I. We show that the proposed approach is guaranteed to asymptotically recover the true graph as if unobserved changing factors were known.

3.1 Assumptions

In this paper, we allow changes in causal modules and some of the changes to be related; the related changes can be explained by positing particular types of confounders. Intuitively, such confounders may refer to some high-level background variables. For instance, for fMRI data, they may be the subject's attention or some unmeasured background stimuli; for the stock market, they may be related to economic policies. Thus, we do not assume causal sufficiency for the set of observed variables. Instead, we assume *pseudo causal sufficiency* as stated below.

Assumption 1 (Pseudo Causal Sufficiency) *We assume that the confounders, if any, can be written as functions of the domain index or smooth functions of time* [2]*. It follows that in each domain or at each time instance, the values of these confounders are fixed.*

To clearly express our basic idea in the presence of distribution shift, we focus on DAGs and assume *pseudo causal sufficiency*. Note that our approach is flexible enough to be extended to cover other types of graphs, e.g., graphs with confounders and graphs with cycles. Later in Section 6.2, we will discuss how nonstationarity helps when there exist stationary confounders. In table 1, we summarize descriptions of different types of confounders (latent common causes) that will be used in this paper, including pseudo confounders, stationary confounders, and nonstationary confounders.

We start with contemporaneous causal relations; the mechanisms and parameters associated with the causal model are allowed to change across data sets or over time, or even vanish or appear in some domains or over some time periods. However, it is natural to generalize our framework to incorporate time-delayed causal relations (Section 6.1).

Denote by $\{g_l(C)\}_{l=1}^{L}$ the set of pseudo confounders (which may be empty). We further assume that for each V_i, its local causal process can be represented by the following structural equation model (SEM):

$$V_i = f_i\big(\mathrm{PA}^i, \mathbf{g}^i(C), \theta_i(C), \epsilon_i\big), \tag{2}$$

2. More specifically, for data with multiple domains, we require that the confounders can be written as a function of the domain index (i.e., it does not change within a domain); for nonstationary time series, we require that the confounder is a smooth function of the time index. Roughly speaking, the smoothness constraint requires the gradient of the function to not change rapidly. In practice, one may specify the level of smoothness in advance (say, by assuming the function follows a Gaussian process prior and properly setting the kernel width to some range) or learn it from data by maximizing marginal likelihood or cross validation.

Confounder type	Description
Pseudo confounder	It can be represented as functions of domain index or smooth functions of time index.
Stationary confounder	Its distribution is fixed and it cannot be represented as functions of domain index or smooth functions of time.
Nonstationary confounder	Its distribution changes across domains or over time and it cannot be represented as functions of domain index or smooth functions of time index, respectively.

Table 1: Descriptions of different types of confounders (latent common causes).

where $\mathbf{g}^i(C) \subseteq \{g_l(C)\}_{l=1}^L$ denotes the set of confounders that influence V_i (it is an empty set if there is no confounder behind V_i and any other variable), $\theta_i(C)$ denotes the effective parameters in the model that are also assumed to be functions of C, and ϵ_i is a disturbance term that is independent of C and PA^i and has a non-zero variance (i.e., the model is not deterministic). We also assume that the ϵ_i's are mutually independent. Note that $\theta_i(C)$ is specific to V_i and is independent of $\theta_j(C)$ for $i \neq j$. The variable C can be the domain or time index. In special cases, e.g., the case with multiple domains and all of which have nonstationarity, C has two dimensions: one is the domain index and the other is the time index. The SEM given in Eq. 2 does not have any restrictions on data distributions or functional classes.

In this paper we treat C as a random variable, and so there is a joint distribution over $\mathbf{V} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_i(C)\}_{i=1}^m$. We assume that this distribution is Markov and faithful to the graph resulting from the following additions to G (which, recall, is the causal structure over $\mathbf{V}$): add $\{g_l(C)\}_{l=1}^L \cup \{\theta_i(C)\}_{i=1}^m$ to G, and for each i, add an arrow from each variable in $\mathbf{g}^i(C)$ to V_i and add an arrow from $\theta_i(C)$ to V_i. We refer to this *augmented graph* as G^{aug}. Obviously, G is simply the induced subgraph of G^{aug} over $\mathbf{V}$. Specifically, the assumption is summarized below.

Assumption 2 *The joint distribution over $\mathbf{V} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_i(C)\}_{i=1}^m$ is Markov and faithful to the augmented graph G^{aug}. In addition, there is no selection bias; i.e., the observed data are perfect random samples from the populations implied by the causal model.*

The distribution change across domains or over time can be considered in the following way. In the case when C is the domain index, C follows a uniform distribution over all possible values, and we have a particular way to generate its value: all possible values are generated once, resulting in domain indices. In the case when C is the time index, we take time to be a special random variable which follows a uniform distribution over the considered time period, with the corresponding data points evenly sampled at a certain sampling frequency. Correspondingly, the generating process of nonstationary data can be considered as follows: we generate random values from C, and then we generate data points over $\mathbf{V}$ according to the SEM in (2). The generated data points are then sorted in ascending order according to the values of C. In other words, we observe the distribution $P(\mathbf{V}|C)$, where $P(\mathbf{V}|C)$ may change across different values of C, resulting in non-identical distributions of data.

3.2 Detection of Changing Modules and Recovery of Causal Skeleton

In this section, we propose a method to detect variables whose causal modules change and infer the skeleton of G. The basic idea is simple: we use the (observed) variable C as a surrogate for the unobserved $\{g_l(C)\}_{l=1}^{L} \cup \{\theta_i(C)\}_{i=1}^{m}$, or in other words, we take C to capture *C-specific* information. We now show that given the assumptions in Section 3.1, we can apply conditional independence tests to $\mathbf{V} \cup C$ to detect variables with changing modules and recover the skeleton of G. We consider C as a surrogate variable (it itself is not a causal variable, it is always available, and confounders and changing parameters are its functions): by adding only C to the variable set $\mathbf{V}$, the skeleton of G and the changing causal modules can be estimated *as if* $\{g_l(C)\}_{l=1}^{L} \cup \{\theta_i(C)\}_{i=1}^{m}$ were known. This is achieved by Algorithm 1 and supported by Theorem 1.

Algorithm 1 Detection of Changing Modules and Recovery of Causal Skeleton

1. Build a complete undirected graph $U_{\mathcal{G}}$ on the variable set $\mathbf{V} \cup C$.
2. (*Detection of changing modules*) For each i, test for the marginal and conditional independence between V_i and C. If they are independent given a subset of $\{V_k \,|\, k \neq i\}$, remove the edge between V_i and C in $U_{\mathcal{G}}$.
3. (*Recovery of causal skeleton*) For every $i \neq j$, test for the marginal and conditional independence between V_i and V_j. If they are independent given a subset of $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$, remove the edge between V_i and V_j in $U_{\mathcal{G}}$.

The procedure given in Algorithm 1 outputs an undirected graph $U_{\mathcal{G}}$ that contains C as well as $\mathbf{V}$. In Step 2, whether a variable V_i has a changing module is decided by whether V_i and C are independent conditional on some subset of other variables. The justification for one side of this decision is trivial. If V_i's module does not change, that means $P(V_i \,|\, \mathrm{PA}^i)$ remains the same for every value of C, and so $V_i \perp\!\!\!\perp C \,|\, \mathrm{PA}^i$. Thus, if V_i and C are not independent conditional on any subset of other variables, V_i's module changes with C, which is represented by an edge between V_i and C. Conversely, we assume that if V_i's module changes, which entails that V_i and C are not independent given PA^i, then V_i and C are not independent given any other subset of $\mathbf{V} \backslash \{V_i\}$. If this assumption does not hold, then we only claim to detect some (but not necessarily all) variables with changing modules.

Step 3 aims to discover the skeleton of the causal structure over $\mathbf{V}$. It leverages the results from Step 2: if neither V_i nor V_j is adjacent to C, then C does not need to be involved in the conditioning set. In practice, one may apply any constraint-based search procedures on $\mathbf{V} \cup C$, e.g., SGS and PC (Spirtes et al., 1993). Its (asymptotic) correctness is justified by the following theorem:

Theorem 1 *Given Assumptions 1 and 2, for every $V_i, V_j \in \mathbf{V}$, V_i and V_j are not adjacent in G if and only if they are independent conditional on some subset of $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$.*

Basic idea of the proof. For a complete proof see Appendix A. The "only if" direction is proved by making use of the weak union property of conditional independence repeatedly, the fact that all $g_l(C)$ and $\theta_i(C)$ are deterministic functions of C, some implications of the

SEMs Eq. 2, the assumptions in Section 3.1, and the properties of mutual information given in Madiman (2008). The "if" direction is shown based on the faithfulness assumption on G^{aug} and the fact that $\{g_l(C)\}_{l=1}^{L} \cup \{\theta_i(C)\}_{i=1}^{m}$ is a deterministic function of C. □

Furthermore, for any pair of nonadjacent variables V_i and V_j with $V_i - C$ and $V_j - C$, we can easily detect whether there are pseudo confounders behind V_i and V_j from the independence test results derived from Algorithm 1:

1. If $\forall \mathbf{V}_k \subseteq \mathbf{V}\backslash\{V_i, V_j\}$, $V_i \not\perp\!\!\!\perp V_j | \mathbf{V}_k$, and $\exists \mathbf{V}_{k'} \subseteq \mathbf{V}\backslash\{V_i, V_j\}$, so that $V_i \perp\!\!\!\perp V_j | \{\mathbf{V}_{k'}, C\}$, then there exist pseudo confounders behind V_i and V_j.

2. If $\exists \mathbf{V}_k \subseteq \mathbf{V}\backslash\{V_i, V_j\}$, so that $V_i \perp\!\!\!\perp V_j | \mathbf{V}_k$, then there is no pseudo confounder behind V_i and V_j.

Note that in Algorithm 1, it is crucial to use a general, nonparametric conditional independence test, for how variables depending on C is unknown and usually very nonlinear. In this work, we use the kernel-based conditional independence test (KCI-test, Zhang et al., 2011) to capture the dependence on C in a nonparametric way. By contrast, if we use, for example, tests of vanishing partial correlations, as is widely used in the neuroscience community, the proposed method may not work well.

Moreover, it is worth noting that the estimated graphical representation by Algorithm 1 naturally follows the principle of minimal changes, which was explicitly formulated by Ghassami et al. (2018). This is because faithfulness on the augmented graph implies the edge minimality condition in the graphical representation. Any variable adjacent to C in the augmented graph has a changing mechanism, and the estimated graph by Algorithm 1 has as few edges involving C as possible; hence it has the smallest number of changing causal mechanisms (or conditional distributions).

4. CD-NOD Phase II: Distribution Shifts Benefit Causal Direction Determination

We now show that introducing the additional variable C as a surrogate not only allows us to infer the skeleton of the causal structure but also facilitates the determination of some causal directions. Let us call those variables that are adjacent to C in the output of Algorithm 1 "C-specific variables", which are actually the effects of changing causal modules. For each C-specific variable V_k, it is possible to determine the direction of every edge which has an endpoint on V_k. Let V_l be any variable adjacent to V_k in the output of Algorithm 1. Then there are two possible scenarios to consider:

S_1. V_l is not adjacent to C. Then $C - V_k - V_l$ forms an unshielded triple. For practical purposes, we take the direction between C and V_k as $C \to V_k$ (though we do not claim C to be a cause in any substantial sense). Then we can use standard orientation rules for unshielded triples to orient the edge between V_k and V_l (Spirtes et al., 1993; Pearl, 2000). There are two possible situations:
1.a If V_l and C are independent given a set of variables excluding V_k, then the triple is a V-structure, and we have $V_k \leftarrow V_l$.
1.b Otherwise, if V_l and C are independent given a set of variables including V_k, then the triple is not a V-structure, and we have $V_k \to V_l$.

S_2. V_l is also adjacent to C. This case is more complex than S_1, but it is still possible to identify the causal direction between V_k and V_l, based on the principle that $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ change independently; a heuristic method is given in Section 4.2.

The procedure in S_1, which will be further discussed in Section 4.1, contains the methods proposed in Hoover (1990); Tian and Pearl (2001); Peters et al. (2016) for causal discovery from changes as special cases. It may also be interpreted as special cases of the principle underlying the method for S_2: if one of $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ changes while the other remains invariant, they are clearly independent.

4.1 Causal Direction Identification by Generalization of Invariance

There exist methods for causal discovery from differences among multiple data sets (Hoover, 1990; Tian and Pearl, 2001; Peters et al., 2016) that explore *invariance* of causal mechanisms. They used linear models to represent causal mechanisms and, as a consequence, the invariance of causal mechanisms can be assessed by checking whether the involved parameters change across data sets or not. Actually, $S_{1.b}$ above provides a nonparametric way to achieve this in light of nonparametric conditional independence tests. For any variable V_i and a set of variables $\mathbf{S}$, the conditional distribution $P(V_i \mid \mathbf{S})$ is invariant across different values of C if and only if

$$P(V_i \mid \mathbf{S}, C = c_1) = P(V_i \mid \mathbf{S}, C = c_2),\ \forall\ c_1 \text{ and } c_2.$$

This is exactly the condition under which $V_i \perp\!\!\!\perp C \mid \mathbf{S}$. In other words, testing for invariance (or homogeneity) of the conditional distribution is naturally achieved by performing a conditional independence test on V_i and C given the set of variables $\mathbf{S}$, for which there exist off-the-shelf algorithms and implementations. When $\mathbf{S}$ is the empty set, this reduces to a test of marginal independence between V_i and C, or a test of homogeneity of $P(V_i)$.

In $S_{1.a}$, we have the invariance of $P(V_l)$ (i.e., $P(\texttt{cause})$) when the causal mechanism, represented by $P(V_k|V_l)$ (i.e., $P(\texttt{effect} \mid \texttt{cause})$), changes, which is complementary to the invariance of causal mechanisms in $S_{1.b}$. The (conditional) independence test results between V_i and C are readily available from Algorithm 1 and can be applied to determine causal directions between variables which satisfy S_1. The procedure is summarized in Algorithm 2.

Algorithm 2 Causal Direction Identification by Generalization of Invariance

1. **Input:** causal skeleton $U_{\mathcal{G}}$ from Algorithm 1.
2. Orient $C \to V_k$, for any variable which is adjacent to C.
3. For any unshielded triple with $C \to V_k - V_l$, where V_l is not adjacent to C,
   a. if $V_l \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $\mathbf{S} \wedge V_k = \emptyset$, orient $V_k \leftarrow V_l$;
   b. if $V_l \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $V_k \in \mathbf{S}$, orient $V_k \to V_l$.
4. **Output:** partially oriented graph $U_{\mathcal{G}}$ using the property of generalization of invariance.

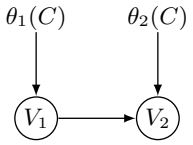


Figure 3: An illustration of a two-variable case: $V_1 \to V_2$ with corresponding parameters $\theta_1(C)$ and $\theta_2(C)$ changing independently.

Naturally, both invariance properties above are particular cases of the principle of *independent changes* of causal modules underlying the method for S_2: if one of $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ changes while the other remains invariant, they are clearly independent. Usually, there is no reason why only one of them could change, so the above invariance properties are rather restrictive. The property of *independent changes* holds in rather generic situations, e.g., when there is no confounder behind `cause` and `effect`. Below we will propose an algorithm for causal direction determination based on independent changes of causal modules.

4.2 Causal Direction Identification by Independently Changing Modules

We now develop a method to handle S_2 above. To clearly express the idea, let us start with a two-variable case: suppose V_1 and V_2 are adjacent and are both adjacent to C. We aim to identify the causal direction between them, which, without loss of generality, we assume to be $V_1 \to V_2$.

Figure 3 shows the case where the involved changing parameters, $\theta_1(C)$ and $\theta_2(C)$, are independent, i.e., $P(V_1; \theta_1)$ and $P(V_2 \mid V_1; \theta_2)$ change independently (we dropped the argument C in θ_1 and θ_2 to simplify notation).

For the reverse direction, one can decompose the joint distribution of (V_1, V_2) according to

$$P(V_1, V_2; \theta_1', \theta_2') = P(V_2; \theta_2')P(V_1 \mid V_2; \theta_1'), \tag{3}$$

where θ_1' and θ_2' are assumed to be sufficient for the corresponding distribution modules $P(V_2)$ and $P(V_1|V_2)$. Generally speaking, θ_1' and θ_2' are not independent, because they are determined jointly by θ_1 and θ_2.

Now we face the problem of how to compare the dependence between θ_1 and θ_2 with that between θ_1' and θ_2'. Since θ is assumed to be sufficient for the corresponding distribution module, it is equivalent to compare the dependence between $P(V_1)$ and $P(V_2|V_1)$ with that between $P(V_2)$ and $P(V_1|V_2)$.

The idea that causal modules are independent is not new (Pearl, 2000), but note that in a stationary situation where each module is fixed, such independence is very difficult, if not impossible, to test. By contrast, in the situation we are considering presently, both $P(V_1)$ and $P(V_2|V_1)$ are changing, and we can try to measure the extent to which variation in $P(V_1)$ and variation in $P(V_2)$ are dependent (and similarly for $P(V_2)$ and $P(V_1|V_2)$). This is the sense in which distribution change actually helps in the identification of causal directions,

and as far as we know, this is the first time that such an advantage is exploited in the case where both $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ change.

We extend the Hilbert Schmidt Independence Criterion (HSIC, Gretton et al., 2008) to measure the dependence between causal modules. To do so, we first develop a novel kernel embedding of nonstationary conditional distributions which does not rely on sliding windows and estimate their corresponding Gram matrices in Section 4.2.1, which will be used in the extended HSIC and the causal direction determination rule in Section 4.2.2. In Section 4.2.3, we propose an algorithm for causal direction determination in multi-variable cases by taking advantage of independent changes.

4.2.1 Kernel Embedding of Constructed Joint Distributions

Notation Throughout this section, we use the following notation. Let X be a random variable on domain $\mathcal{X}$, and $(\mathcal{H}, k)$ be a Reproducing Kernel Hilbert Space (RKHS) with a measurable kernel on $\mathcal{X}$. Let $\phi(x) \in \mathcal{H}$ represent the feature map for each $x \in \mathcal{X}$, with $\phi : \mathcal{X} \to \mathcal{H}$. We assume integrability: $E_X[k(X, X)] \leq \infty$. Similar notations are for variables Y and C. The cross-covariance operator $C_{YX} : \mathcal{H} \to \mathcal{G}$ is defined as $C_{YX} := E_{YX}[\phi(X) \otimes \psi(Y)]$, where $\mathcal{G}$ is the RKHS associated with Y.

We represent causal modules $P(V_i|\mathrm{PA}^i, C)$ by kernel embedding. Intuitively, to represent the kernel embedding of changing causal modules, we need to consider $P(V_i|\mathrm{PA}^i, C)$ for each value of C separately. If C is a domain index, for each value of C we have a dataset of (V_i, PA^i). If C is a time index, one may use a sliding window to use the data of (V_i, PA^i) in the window of length L centered at $C = c$. However, in some cases, it might be hard to find an appropriate window length L, especially when the causal module changes fast. In the following, we propose a way to estimate the kernel embedding of changing causal modules on the whole dataset, avoiding window segmentation. For the sake of conciseness, below we use Y and X to denote V_i and PA^i, respectively.

Suppose that there are N samples for each variable. Instead of working with $P(Y|X, C = c_n)$ $(n = 1, \cdots, N)$ directly, we "virtually" construct a particular distribution $\tilde{P}(\underline{Y}, X \mid C = c_n)$ as follows:[3]

$$\tilde{P}(\underline{Y}, X|C = c_n) = P(Y|X, C = c_n)P(X). \tag{4}$$

The embedding of this "joint distribution" of X and Y is simpler than that of the conditional of Y given X. Since $P(X)$ does not depend on C and its support is rich enough to contain that of $P(X|C = c_n)$, one can see that whenever there are changes in $P(Y|X, C = c_n)$ across different values of c_n, there must be changes in $\tilde{P}(\underline{Y}, X|C = c_n)$, and vice versa. In other words, the constructed distribution $\tilde{P}(\underline{Y}, X|C = c_n)$ captures changes in $P(Y|X, C = c_n)$ across different c_n. We let $\tilde{P}(\underline{Y}, X, C = c_n) = P(Y|X, C = c_n)P(X)P(C = c_n)$.

Proposition 1 shows that the kernel embedding of the distribution $\tilde{P}(\underline{Y}, X|C = c_n)$ can be estimated on the whole data set, without window segmentation.

Proposition 1 *Let X represent the direct causes of Y, and suppose that they have N observations. The kernel embedding of distribution $\tilde{P}(\underline{Y}, X|C = c_n)$ can be represented as*

$$\hat{\tilde{\mu}}_{\underline{Y},X|C=c_n} = \frac{1}{n}\mathbf{\Phi_y}(\mathbf{K_x} \odot \mathbf{K_c} + \lambda I)^{-1} diag(\mathbf{k}_{\mathbf{c},c_n})\mathbf{K_x}\mathbf{\Phi}_{\mathbf{x}}^{\intercal},$$

3. Here we use $\underline{Y}$ instead of Y to emphasize that in this constructed distribution Y and X are not symmetric.

where $\mathbf{\Phi_y} := [\phi(y_1), ..., \phi(y_N)]$, $\mathbf{\Phi_x} := [\phi(x_1), ..., \phi(x_N)]$, $\mathbf{k_{c,c_n}} := [k(c_1, c_n), ..., k(c_N, c_n)]^\intercal$, $\mathbf{K_x}(x_t, x_{t'}) = \langle \phi(x_t), \phi(x_{t'}) \rangle$, $\mathbf{K_c}(c_t, c_{t'}) = \langle \phi(c_t), \phi(c_{t'}) \rangle$, *and* $\odot$ *represents point-wise product.*

The detailed proof of proposition 1 is given in Appendix B. Next we estimate the $N \times N$ Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C=c}$. We consider different kernels for the estimation of Gram matrix. Let $M^l_{\underline{Y}X}$ represent the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C=c}$ with a linear kernel, and $M^{\mathcal{G}}_{\underline{Y}X}$ the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C=c}$ with a Gaussian kernel.

If we use a linear kernel, the (c, c')th entry of the Gram matrix $M^l_{\underline{Y}X}$ is the inner product between $\hat{\tilde{\mu}}_{\underline{Y},X|C=c}$ and $\hat{\tilde{\mu}}_{\underline{Y},X|C=c'}$:

$$\begin{aligned} M^l_{\underline{Y}X}(c, c') \triangleq & \mathrm{tr}(\hat{\tilde{\mu}}^\intercal_{\underline{Y},X|C=c} \hat{\tilde{\mu}}_{\underline{Y},X|C=c'}) \\ = & \frac{1}{N^2} \mathbf{k}^\intercal_{\mathbf{c},c} \big[\mathbf{K}^3_{\mathbf{x}} \odot \big((\mathbf{K_x} \odot \mathbf{K_c} + \lambda I)^{-1} \mathbf{K_y} (\mathbf{K_x} \odot \mathbf{K_c} + \lambda I)^{-1} \big) \big] \mathbf{k}_{\mathbf{c},c'}, \end{aligned} \tag{5}$$

which is the (c, c')th entry of the matrix

$$\mathbf{M}^l_{\underline{Y}X} = \frac{1}{N^2} \mathbf{K_c} \big[\mathbf{K}^3_{\mathbf{x}} \odot \big((\mathbf{K_x} \odot \mathbf{K_c} + \lambda I)^{-1} \mathbf{K_y} (\mathbf{K_x} \odot \mathbf{K_c} + \lambda I)^{-1} \big) \big] \mathbf{K_c}. \tag{6}$$

If we use a Gaussian kernel with kernel width σ_2, the Gram matrix is given by

$$\begin{aligned} M^{\mathcal{G}}_{\underline{Y}X}(c, c') = & \exp\big(- \frac{||\tilde{\mu}_{\underline{Y},X|C=c} - \tilde{\mu}_{\underline{Y},X|C=c'}||^2_F}{2\sigma_2^2} \big) \\ = & \exp\big(- \frac{M^l_{\underline{Y}X}(c, c) + M^l_{\underline{Y}X}(c', c') - 2M^l_{\underline{Y}X}(c', c)}{2\sigma_2^2} \big), \end{aligned} \tag{7}$$

where $|| \cdot ||_F$ denotes the Frobenius norm. This can be represented in matrix notation as

$$\mathbf{M}^{\mathcal{G}}_{\underline{Y}X} = \exp\big(- \frac{diag(\mathbf{M}^l_{\underline{Y}X}) \cdot \mathbf{1}_N + \mathbf{1}_N \cdot diag(\mathbf{M}^l_{\underline{Y}X}) - 2\mathbf{M}^l_{\underline{Y}X}}{2\sigma_2^2} \big), \tag{8}$$

where $diag(\cdot)$ sets all off-diagonal entries zero, and $\mathbf{1}_N$ is an $N \times N$ matrix with all entries being 1.

Excitingly, we can see that with our methods, we do not need to explicitly learn the high-dimensional kernel embedding $\tilde{\mu}_{\underline{Y},X|C=c}$ for each c. With the kernel trick, the final Gram matrix can be represented by $N \times N$ kernel matrices directly.

There are several hyperparameters to set. The hyperparameters associated with $\mathbf{K_x}$, $\mathbf{K_c}$, and the regularization parameter λ in equation (6) are learned through a Gaussian process regression framework: they are learned by maximizing the marginal likelihood of Y. For the hyperparameters associated with $\mathbf{K_y}$ and the kernel with σ_2 in equation (7), we set them with empirical values; please refer to Zhang et al. (2011) for details.

Change in marginal distributions. As a special case, when we are concerned with how the marginal distribution of Y changes with C, i.e., when $X = \emptyset$, we directly make use of

$$\mu_{Y|C=c_n} = \mathcal{C}_{YC} \mathcal{C}^{-1}_{CC} \phi(c_n). \tag{9}$$

This can also be obtained by constraining X in $\tilde{\mu}_{\underline{Y},X|C=c_n}$ to take a fixed value. Its empirical estimate is

$$\begin{aligned}\hat{\mu}_{Y|C=c_n} &= \frac{1}{N}\mathbf{\Phi_y}\mathbf{\Phi_c}^\intercal(\frac{1}{n}\mathbf{\Phi_c}\mathbf{\Phi_c}^\intercal + \lambda I)^{-1}\phi_{c_n} \\ &= \mathbf{\Phi_y}(\mathbf{K_c} + \lambda I)^{-1}\mathbf{k}_{\mathbf{c},c_n}. \end{aligned} \tag{10}$$

Then (c, c') entry of the Gram matrix with a linear kernel is:

$$\begin{aligned} M_Y^l(c,c') &\triangleq \hat{\mu}_{Y|C=c}^\intercal \hat{\mu}_{Y|C=c'} \\ &= \mathbf{k}_{\mathbf{c},c}^\intercal(\mathbf{K_c} + \lambda I)^{-1}\mathbf{\Phi_y}^\intercal\mathbf{\Phi_y}(\mathbf{K_c} + \lambda I)^{-1}\mathbf{k}_{\mathbf{c},c'} \\ &= \mathbf{k}_{\mathbf{c},c}^\intercal(\mathbf{K_c} + \lambda I)^{-1}\mathbf{K_y}(\mathbf{K_c} + \lambda I)^{-1}\mathbf{k}_{\mathbf{c},c'}, \end{aligned} \tag{11}$$

which is the (c, c')th entry of

$$\mathbf{M}_Y^l = \mathbf{K_c}(\mathbf{K_c} + \lambda I)^{-1}\mathbf{K_y}(\mathbf{K_c} + \lambda I)^{-1}\mathbf{K_c}. \tag{12}$$

For a Gaussian kernel with kernel with σ_2, the Gram matrix is

$$\mathbf{M}_Y^{\mathcal{G}} = \exp\big(-\frac{diag(\mathbf{M}_{\underline{Y}X}^l)\cdot\mathbf{1}_N + \mathbf{1}_N\cdot diag(\mathbf{M}_{\underline{Y}X}^l) - 2\mathbf{M}_{\underline{Y}X}^l}{2\sigma_2^2}\big). \tag{13}$$

4.2.2 Two-Variable Case

In this section, we extend HSIC to measure the dependence between causal modules, based on which we determine causal directions.

For simplicity, let us start with the two-variable case: suppose that X and Y are adjacent and both are adjacent to C. We aim to identify the causal direction between them, which, without loss of generality, we assume to be $X \to Y$. The guiding idea is that distribution shift may carry information that confirms the independence of causal modules, which, in the simple case we are considering, is the independence between $P(X)$ and $P(Y|X)$. If $P(X)$ and $P(Y|X)$ are independent but $P(Y)$ and $P(X|Y)$ are not, then the causal direction is inferred to be from X to Y.

The dependence between $P(X)$ and $P(Y|X)$ can be measured by extending the HSIC (Gretton et al., 2008).

HSIC Given a set of observations $\{(u_1, v_1), (u_2, v_2), ..., (u_N, v_N)\}$ for variables U and V, HSIC provides a measure of dependence and a statistic for testing their statistical independence. Roughly speaking, it measures the squared covariances between feature maps of U and feature maps of V. Let $\mathbf{M}_U$ and $\mathbf{M}_V$ be the Gram matrices for U and V calculated on the sample, respectively. An estimator of HSIC is given by Gretton et al. (2008):

$$\text{HSIC}_{UV} = \frac{1}{(N-1)^2}\mathtt{tr}(\mathbf{M}_U H \mathbf{M}_V H), \tag{14}$$

where H is used to center the features, with entries $H_{ij} := \delta_{ij} - N^{-1}$.

In what follows, we will use a normalized version of the estimated HSIC, which is invariant to the scale in $\mathbf{M}_U$ and $\mathbf{M}_V$:

$$\begin{aligned}\text{HSIC}_{UV}^{\mathcal{N}} &= \frac{\text{HSIC}_{UV}}{\frac{1}{N-1}\texttt{tr}(\mathbf{M}_U H)\cdot\frac{1}{N-1}\texttt{tr}(\mathbf{M}_V H)} \\ &= \frac{\texttt{tr}(\mathbf{M}_U H \mathbf{M}_V H)}{\texttt{tr}(\mathbf{M}_U H)\texttt{tr}(\mathbf{M}_V H)}.\end{aligned} \tag{15}$$

Dependence between Changing Modules In our case, we aim to check whether $P(Y|X)$ and $P(X)$ change independently along with C. We work with the estimate of their embeddings. Then we can think of $\left\{\left(\hat{\mu}_{X|C=c_n}, \hat{\tilde{\mu}}_{\underline{Y},X|C=c_n}\right)\right\}_{n=1}^{N}$ as the observed data pairs and measure their dependence from the data pairs.

This can be done by applying (the normalized version of) the estimate of HSIC given in Eq. 15 to the above data pairs. The expression then involves $\mathbf{M}_X$, the Gram matrix of $\hat{\mu}_{X|C}$ at $C = c_1, c_2, ..., c_N$, and $\mathbf{M}_{\underline{Y}X}$, the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y}X|C}$ at $C = c_1, c_2, ..., c_N$. In particular, the dependence between $P(Y|X)$ and $P(X)$ on the given data can be estimated by

$$\hat{\Delta}_{X\to Y} = \frac{\texttt{tr}(\mathbf{M}_X H \mathbf{M}_{\underline{Y}X} H)}{\texttt{tr}(\mathbf{M}_X H)\texttt{tr}(\mathbf{M}_{\underline{Y}X} H)}. \tag{16}$$

Similarly, for the hypothetical direction $Y \to X$ the dependence between $P(X|Y)$ and $P(Y)$ on the data is estimated by

$$\hat{\Delta}_{Y\to X} = \frac{\texttt{tr}(\mathbf{M}_Y H \mathbf{M}_{\underline{X}Y} H)}{\texttt{tr}(\mathbf{M}_Y H)\texttt{tr}(\mathbf{M}_{\underline{X}Y} H)}. \tag{17}$$

We then have the following rule to determine the causal direction between X and Y.

Causal Direction Determination Rule Suppose that X and Y are two random variables with N observations. We assume that X and Y are adjacent and both are adjacent to C and assume no pseudo confounders behind them. The causal direction between X and Y is then determined according to the following rule:

- if $\hat{\Delta}_{X\to Y} < \hat{\Delta}_{Y\to X}$, output $X \to Y$;
- if $\hat{\Delta}_{X\to Y} > \hat{\Delta}_{Y\to X}$, output $X \leftarrow Y$.

In practice, there may exist pseudo confounders. In such a case, we set a threshold α on $\hat{\Delta}$. If $\hat{\Delta}_{X\to Y} > \alpha$ and $\hat{\Delta}_{Y\to X} > \alpha$, we conclude that there are pseudo confounders which influence both X and Y and leave the direction undetermined.

4.2.3 With More Than Two Variables

The causal direction determination rule in the two-variable case can be extended to learn causal directions in multi-variable cases. Suppose that we have m observed random variables $\{V_i\}_{i=1}^m$ and a partially oriented graph $U_{\mathcal{G}}$ derived from Algorithms 1 and 2. Let $\mathbf{V}_S$ be the subset of $\{V_i\}_{i=1}^m$, such that $V_i \in \mathbf{V}_S$ if and only if V_i's causal module changes.

Note that generally speaking, different from the unconfounded two-variable case, when identifying the causal direction between an unoriented pair of adjacent variables, we need to remove the effect from their common causes. Thus, before moving forward, we first define *deconfounding set* and *potential deconfounding set* of a pair of adjacent variables in $U_{\mathcal{G}}$.

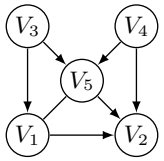


Figure 4: An illustration of the definitions of minimal deconfounding set and minimal potential deconfounding set on a partially oriented graph.

Definition 1 (Deconfounding Set) *A set of variables $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_l, V_k\}$ is the deconfounding set of a pair of adjacent variables (V_l, V_k), if*

(i) no node in $\mathbf{Z}$ is a descendant of V_l or V_k,

(ii) and $\mathbf{Z}$ blocks every path between V_l and V_k that contains arrows into V_l and V_k.

Furthermore, a set of variables $\mathbf{Z}$ is the *minimal deconfounding set* of a pair of adjacent variables (V_l, V_k), if any $\mathbf{Z}_s \subset \mathbf{Z}$ is not a deconfounding set.

Definition 2 (Potential Deconfounding Set) *A set of variables $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_l, V_k\}$ is the potential deconfounding set of a pair of adjacent variables (V_l, V_k), if*

(i) no node in $\mathbf{Z}$ is a descendant of V_l or V_k,

(ii) $\mathbf{Z}$ blocks every path between V_l and V_k that does not contain an arrow out of V_l or V_k,

(iii) and any $Z \in \mathbf{Z}$ is not in the deconfounding set.

Similarly, a set of variables $\mathbf{Z}$ is the *minimal potential deconfounding set* of a pair of adjacent variables (V_l, V_k), if any subset $\mathbf{Z}_s \subset \mathbf{Z}$ is not a potential deconfounding set.

In Figure 4, for example, the set $\mathbf{Z} = \{V_3\}$ is a minimal deconfounding set of (V_1, V_2), and the set $\mathbf{Z} = \{V_5\}$ is a minimal potential deconfounding set of (V_1, V_2).

We take advantage of the independence between causal modules to identify directions. To efficiently identify causal directions using independent changes when there are multiple variables, we propose Algorithm 3, with the main procedure as follows. For each undirected pair $V_l, V_k \in \mathbf{V}_S$, we denote their minimal deconfounding set by $\mathbf{Z}_{lk}^{(1)}$ and their minimal potential deconfounding set by $\mathbf{Z}_{lk}^{(2)}$. Note that there may be multiple minimal deconfounding sets and multiple minimal potential deconfounding sets; to search for such sets more efficiently, in our implementation, we only consider those ones with every variable in $\mathbf{Z}_{lk}^{(1)}$ and $\mathbf{Z}_{lk}^{(2)}$ being adjacent to V_l. Let $\mathbf{Z}_{lk}^{(2,n)}$ be a subset of $\mathbf{Z}_{lk}^{(2)}$, where n is the total cardinality of $\mathbf{Z}_{lk}^{(1)}$ and $\mathbf{Z}_{lk}^{(2,n)}$; i.e., $|\mathbf{Z}_{lk}^{(1)}| + |\mathbf{Z}_{lk}^{(2,n)}| = n$. We evaluate the dependence between $P(V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and $P(V_l | V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$, and that between $P(V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and $P(V_k | V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$. If we find that $P(V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \perp\!\!\!\perp P(V_k | V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and that $P(V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \not\perp\!\!\!\perp P(V_l | V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$, we output $V_l \to V_k$, and if there are unoriented edges from variables in $\mathbf{Z}_{lk}^{(2,n)}$ to V_k or V_l, then we consider those variables as

parents. Similarly, if we find that $P(V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \perp\!\!\!\perp P(V_l|V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and that $P(V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \not\perp\!\!\!\perp P(V_k|V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$, we output $V_k \to V_l$, instead. Similar to the search procedure of PC, we start from $n = 0$, evaluate the dependence between corresponding modules for each undirected pair, and then let $n = n + 1$ and repeat the procedure until no unoriented pairs have the total cardinality of minimal deconfounding set and minimal potential deconfounding set greater than or equal to n.

Note that in Algorithm 3, we use dependence measures in (16) and (17) to determine the independence between causal modules. Particularly, we set a threshold α. If the dependence measure $\hat{\Delta} \leq \alpha$, then the corresponding modules are independent; otherwise, they are dependent. For a pair of adjacent variables $V_k, V_l \in \mathbf{V}_S$, if their direction is undetermined by Algorithm 3 and all their measured modules are dependent, then there exist pseudo confounders behind V_k and V_l.

Furthermore, for variables whose causal modules are stationary and which are adjacent only to variables with stationary modules, the causal directions between them cannot be determined by Algorithm 2 or 3. In such a case, one may further infer some causal directions by making use of Meek's orientation rules (Meek, 1995) used in PC.

To better illustrate the algorithm, we go through a simple example, for which the true graph and brief orientation procedures are given in Figure 5. Below is the precise orientation procedure.

1. In step 1, we have observed variables $\{V_i\}_{i=1}^4$, with $V_i \in \mathbf{V}_S$ for any i, and an unoriented graph $U_{\mathcal{G}}$ over $\{V_i\}_{i=1}^4$ after Algorithms 1 and 2. Since all causal modules are changing, there is no invariance property that can be used for orientation determination in Algorithm 2.
2. In step 2, we start from $n = 0$. For example, for the unoriented pair of adjacent variables (V_1, V_2), $\mathbf{Z}_{12}^{(1)} = \{\emptyset\}$ and $\mathbf{Z}_{12}^{(2,n)} = \{\emptyset\}$. In this case, we have $P(V_1) \not\perp\!\!\!\perp P(V_2|V_1)$ and $P(V_2) \not\perp\!\!\!\perp P(V_1|V_2)$, and thus we cannot determine the direction between V_1 and V_2. Similarly, we cannot determine the direction between V_2 and V_4. For the unoriented pair (V_1, V_3), we have that $P(V_3) \perp\!\!\!\perp P(V_1|V_3)$ and that $P(V_1) \not\perp\!\!\!\perp P(V_3|V_1)$, and thus we output $V_3 \to V_1$. Similarly, we can determine the direction between V_3 and V_4 and output $V_3 \to V_4$.

   Then let $n = 1$. For the unoriented pair (V_1, V_2), $\mathbf{Z}_{12}^{(1)} = \{\emptyset\}$ and $\mathbf{Z}_{12}^{(2,n)} = \{V_3\}$. We then have that $P(V_1, V_3) \perp\!\!\!\perp P(V_2|V_1, V_3)$ and that $P(V_2, V_3) \not\perp\!\!\!\perp P(V_1|V_2, V_3)$, and thus we output $V_1 \to V_2$. For the unoriented pair (V_2, V_4), $\mathbf{Z}_{24}^{(1)} = \{V_3\}$ and $\mathbf{Z}_{24}^{(2,n)} = \{\emptyset\}$. We then have that $P(V_4, V_3) \perp\!\!\!\perp P(V_2|V_4, V_3)$ and that $P(V_2, V_3) \not\perp\!\!\!\perp P(V_4|V_2, V_3)$, and thus output $V_4 \to V_2$.
3. We output the fully identified causal graph.

4.3 Identifiability Conditions of CD-NOD

In this section, we first give identifiability conditions of CD-NOD, and then we define the equivalence class that CD-NOD can identify if corresponding conditions do not hold.

We make use of independent changes and orientation rules, including discovering V-structures and using orientation propagation, to determine causal directions. To fully identify the DAG, we ought to make some assumption on the change property of the distribution in wrong directions. In particular, we have the following assumption.

Algorithm 3 Causal Direction Identification by Independent Changes of Causal Modules

1. **Input**: observations of $\{V_i\}_{i=1}^m$, a subset $\mathbf{V}_S \subseteq \mathbf{V}$ ($\forall V \in \mathbf{V}_S$, V has a changing causal module), a partially oriented causal graph $U_{\mathcal{G}}$ from Algorithms 1 and 2.

2. $n = 0$
   Repeat

   Repeat

   i. Select an unoriented pair of adjacent variables (V_l, V_k), with $V_l, V_k \in \mathbf{V}_S$.

   A. Let $\mathbf{Z}_{lk}^{(1)}$ be the set of minimal deconfounding set of V_k and V_l; any $Z \in \mathbf{Z}_{lk}^{(1)}$ is adjacent to V_l.

   B. Let $\mathbf{Z}_{lk}^{(2)}$ be the set of minimal potential deconfounding set; any $Z \in \mathbf{Z}_{lk}^{(2)}$ is adjacent to V_l.

   ii. If $|\mathbf{Z}_{lk}^{(1)}| + |\mathbf{Z}_{lk}^{(2)}| \geq n$, repeat

   A. Take $\mathbf{Z}_{lk}^{(2,n)} \subseteq \mathbf{Z}_{lk}^{(2)}$, with $|\mathbf{Z}_{lk}^{(2,n)}| = n - |\mathbf{Z}_{lk}^{(1)}|$.

   B. If $P(V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \perp\!\!\!\perp P(V_l | V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and $P(V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \not\perp\!\!\!\perp P(V_k | V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$, output $V_k \to V_l$;
   for $Z \in \mathbf{Z}_{lk}^{(2,n)}$ with $Z - V_l$, output $Z \to V_l$;
   for $Z \in \mathbf{Z}_{lk}^{(2,n)}$ with $Z - V_k$, output $Z \to V_k$.

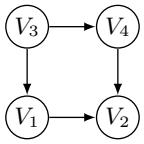
   C. If $P(V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \not\perp\!\!\!\perp P(V_l | V_k, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$ and $P(V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)}) \perp\!\!\!\perp P(V_k | V_l, \mathbf{Z}_{lk}^{(1)}, \mathbf{Z}_{lk}^{(2,n)})$, output $V_l \to V_k$;
   for $Z \in \mathbf{Z}_{lk}^{(2,n)}$ with $Z - V_l$, output $Z \to V_l$;
   for $Z \in \mathbf{Z}_{lk}^{(2,n)}$ with $Z - V_k$, output $Z \to V_k$.

   D. If one of the conditions in (B) and (C) holds, return to step (i);

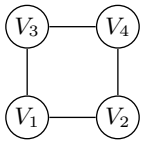
   Until every unoriented pair of adjacent variables (V_l, V_k), with $V_l, V_k \in \mathbf{V}_S$, has been selected.

   $n = n + 1$.

3. **Output**: graph $U_{\mathcal{G}}$, with edges between variables in $\mathbf{V}_S$ oriented.



True graph



Graph after Algorithms 1 and 2

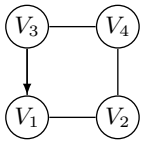
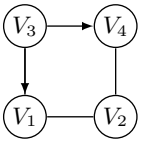
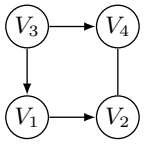
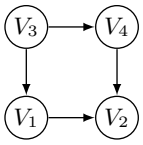
	Independencies & Dependencies	Resulting graphs
$n = 0$:	$P(V_3) \perp\!\!\!\perp P(V_1\|V_3)$ & $P(V_1) \not\perp\!\!\!\perp P(V_3\|V_1)$. So output $V_3 \to V_1$.	
	$P(V_3) \perp\!\!\!\perp P(V_4\|V_3)$ & $P(V_4) \not\perp\!\!\!\perp P(V_3\|V_1)$. So output $V_3 \to V_4$.	
$n = 1$:	$P(V_1, V_3) \perp\!\!\!\perp P(V_2\|V_1, V_3)$ & $P(V_2, V_3) \not\perp\!\!\!\perp P(V_1\|V_2, V_3)$. So output $V_1 \to V_2$.	
	$P(V_4, V_3) \perp\!\!\!\perp P(V_2\|V_4, V_3)$ & $P(V_2, V_3) \not\perp\!\!\!\perp P(V_4\|V_2, V_3)$. So output $V_4 \to V_2$.	

Figure 5: An example to illustrate causal direction identification with Algorithm 3.

Assumption 3 *If $V_i \to V_j$, with $V_i, V_j \in \mathbf{V}$, and at least one of them has a changing mechanism, then $P(V_i|V_j, \mathbf{Z})$ and $P(V_j, \mathbf{Z})$ are dependent, where $\mathbf{Z} \subseteq \mathbf{V}\backslash\{V_l, V_k\}$ is a minimal deconfounding set of (V_i, V_j).*

This assumption can be viewed as faithfulness on another level. Specifically, if $V_i \to V_j$, $P(V_i|V_j, \mathbf{Z})$ and $P(V_j, \mathbf{Z})$ involve the changing parameters both in V_i and V_j, so they are expected to be dependent. Now we are ready to give identifiability conditions of causal directions, which are stated in Theorem 2.

Theorem 2 *Under Assumptions 1, 2, and 3, the causal direction between adjacent variables $V_i, V_j \in \mathbf{V}$ is identifiable with CD-NOD, if at least one of the following three conditions is satisfied:*

(1) the edge between V_i and V_j is involved in a V-structure;

(2) at least one of V_i's causal module and V_j's causal module changes, and if both, the causal modules of V_i and V_j are independent;

(3) there exists an edge incident to one and only one of the variables in $\{V_i, V_j\}$.

Please refer to Appendix C for detailed proofs. Note that CD-NOD does not require hard restrictions on the functional class of causal mechanisms and data distributions. Combined with Theorem 1, we know that for any pair of adjacent variables in graph G, if at least one of the three conditions in Theorem 2 holds, then the whole causal structure is identifiable. It is given in the following corollary.

Corollary 1 *Under Assumptions 1, 2, and 3, the whole causal graph is identifiable, if any pair of adjacent variables V_i and V_j in graph G satisfies at least one of the following three conditions:*

(1) the edge between V_i and V_j is involved in a V-structure;

(2) at least one of V_i's causal module and V_j's causal module changes, and if both, the causal modules of V_i and V_j are independent;

(3) there exists an edge incident to one and only one of the variables in $\{V_i, V_j\}$.

Proof From Theorem 1, we know that the causal skeleton is identifiable. In addition, from Theorem 2, we know that for any pair of adjacent variables, if at least one of the three conditions is satisfied, then the direction is identifiable. Therefore, the whole causal structure is identifiable. ■

For a pair of adjacent variables V_i and V_j, if none of the conditions in Theorem 2 holds, we may not be able to determine the causal direction between them. Thus, in this case, we cannot derive a fully identified causal graph but an equivalence class. Note that the causal skeleton is identifiable, as shown by Theorem 1, regardless of these conditions. Below we give a formal definition of equivalence class with CD-NOD.

Definition 3 *(CD-NOD Equivalence Class) Let $G = (\mathbf{V}, \mathbf{E})$ and $G' = (\mathbf{V}, \mathbf{E}')$ be two DAGs over the same set of variables $\mathbf{V}$. G and G' are called CD-NOD equivalent, if and only if they satisfy the following properties.*

(1) G and G' have the same causal skeleton.

(2) For any pair of adjacent variables $V_i, V_j \in \mathbf{V}$, if it satisfies at least one of the conditions given in Theorem 2, then the causal direction between V_i and V_j is the same in G and G'.

Remark: Another important factor that needs to be considered in practical problems is computational complexity. In CD-NOD phase I, the computational complexity of each kernel-based (conditional) independence test is $O(N^3)$, where N is the number of samples of each variable. The computational complexity of the search procedure with PC is bounded by $\frac{(m+1)^2 m^{k-1}}{(k-1)!}$, where m is the number of observed variables, and k is the maximal degree of variables in $\mathbf{V} \cup C$. In CD-NOD phase II, the computational complexity of each dependence measure, with normalized HSIC, is $O(N^3)$.

5. CD-NOD Phase III: Nonstationary Driving Force Estimation

In this section, we focus on the visualization of *how* causal module $P(V_i \,|\, \mathrm{PA}^i, C)$ changes, i.e., where the changes occur, how fast it changes, and how to visualize the changes. We assume that we already know the causal structure and know which causal modules are changing (see Algorithms 1, 2, and 3).

In the parametric case, if we know which parameters of the causal model $\mathrm{PA}^i \to V_i$ are changing, e.g., the mean of a direct cause, the coefficients in a linear SEM, then we can estimate such parameters for different values of C and see how they change. However, such knowledge is usually not available, and for the sake of flexibility, it is generally better to model causal processes nonparametrically. Therefore, it is desirable to develop a general nonparametric procedure to capture the change of causal modules.

We aim to find a low-dimensional and interpretable mapping of $P(V_i \,|\, \mathrm{PA}^i, C)$ which captures its nonstationarity in a nonparametric way:

$$\lambda_i(C) = h_i(P(V_i \,|\, \mathrm{PA}^i, C)), \tag{18}$$

where h_i is a nonlinear function, mapping the conditioning distribution $P(V_i \,|\, \mathrm{PA}^i, C)$ to a low dimensional representation $\lambda_i(C)$. We call $\lambda_i(C)$ the nonstationary driving force of $P(V_i \,|\, \mathrm{PA}^i, C)$. This formulation is rather general: any identifiable parameters in $P(V_i \,|\, \mathrm{PA}^i, C)$ can be expressed this way, and in the nonparametric case, $\lambda_i(C)$ can be seen as a statistic to summarize changes in $P(V_i \,|\, \mathrm{PA}^i, C)$ along with different values of C. If $P(V_i \,|\, \mathrm{PA}^i, C)$ does not change along with C, then $\lambda_i(C)$ remains constant; otherwise, $\lambda_i(C)$ is intended to capture the variability of $P(V_i \,|\, \mathrm{PA}^i, C)$ across different values of C.

Now there are two problems to solve. One is that given only observed data, how we represent the conditional distributions conveniently. The other is what method to use to enable $\lambda_i(C)$ to capture the variability in the conditional distribution along with C.

For the former problem, we represent changing conditional distributions by kernel embedding as that given in Proposition 1, which is readily achievable. For the latter one,

we use kernel principle component analysis (KPCA) to capture its variability along with C and accordingly propose a method called Kernel Nonstationary Visualization (KNV) of causal modules. In the following, for conciseness, we will use X and Y to denote V_i and PA^i, respectively.

Nonstationary Driving Force Estimation as Eigenvalue Decomposition Problems We use the estimated kernel embedding of distributions, $\hat{\tilde{\mu}}_{\underline{Y},X|C=c_n}$ $(n = 1, \cdots, N)$, derived from Proposition 1 as the input, and aim to find $\hat{\lambda}(C)$ as a (linear or nonlinear) transformation of $\tilde{\mu}_{\underline{Y},X|C=c_n}$, to capture its variability across different values of C. This can be achieved by exploiting KPCA techniques (Schölkopf et al., 1998), which computes principal components in kernel spaces of the input.

To perform KPCA, we need to know the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C=c_n}$, which has already been estimated in Section 4.2.1. We use $M^l_{\underline{Y}X}$ to represent the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C}$ with a linear kernel, and $M^{\mathcal{G}}_{\underline{Y}X}$ the Gram matrix of $\hat{\tilde{\mu}}_{\underline{Y},X|C}$ with a Gaussian kernel. Then $\hat{\lambda}(C)$ can be found by performing eigenvalue decomposition on the above Gram matrix, $M^l_{\underline{Y}X}$ or $M^g_{\underline{Y}X}$; for details please see Schölkopf et al. (1998). In practice, one may take the first few eigenvectors which capture most of the variance.

We can see that with our methods, we do not need to explicitly learn the high-dimensional kernel embedding $\tilde{\mu}_{\underline{Y},X|C=c}$ for each c. With the kernel trick, the final Gram matrix can be represented by $N \times N$ kernel matrices directly. Then the nonstationary driving force $\hat{\lambda}(C)$ can be estimated by performing eigenvalue decomposition on the Gram matrix. Algorithm 4 summarizes the proposed KNV method.

Algorithm 4 KNV of Causal Modules $P(Y|X, C)$

1. **Input:** N observations of X and Y.
2. Calculate Gram matrix $M_{\underline{Y}X}$ (see Eq. 6 for linear kernels and Eq. 8 for Gaussian kernels).
3. Find $\hat{\lambda}(C)$ by directly feeding Gram matrix $M_{\underline{Y}X}$ to KPCA. That is, perform eigenvalue decomposition on $M_{\underline{Y}X}$ to find the nonlinear principal components $\hat{\lambda}(C)$, as in Section 4.1 of Schölkopf et al. (1998). In practice, we may take the first few eigenvectors which capture most of the variance.
4. **Output:** the estimation of nonstationary driving force $\hat{\lambda}(C)$ of $P(Y|X, C)$.

6. Extensions of CD-NOD

In this section, we show how to extend CD-NOD to deal with more general scenarios. For example, we extend CD-NOD to allow both time-varying instantaneous and lagged causal relationships. We additionally discuss whether and how distribution shifts help for causal discovery in the presence of stationary confounders. We then give a procedure, which

leverages both CD-NOD and approaches based on constrained functional causal models, to extend the generality of the proposed method.

6.1 With Time-Varying Instantaneous and Lagged Causal Relationships

In many scenarios, e.g., dynamic systems with insufficient time resolution, there may exist both instantaneous and time-lagged causal relationships over the measured data (Hyvärinen et al., 2010; Gong* et al., 2015). In this section, we extend CD-NOD proposed in Section 3 to recover both time-varying instantaneous and lagged causal relationships and identify changing causal modules.

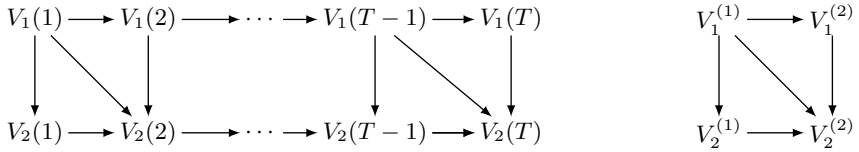
Suppose that there are m observed processes $\mathbf{V}(t) = (V_1(t), \cdots, V_m(t))^{\mathrm{T}}$ with $t = 1, \cdots, T$, where there exist both time-varying instantaneous and lagged causal relations over these m processes. Further denote the largest time lag by P. To efficiently recover both instantaneous and time-lagged causal relations, we reorganize the set of variables $\mathbf{V}$ into $\tilde{\mathbf{V}} = \tilde{\mathbf{V}}^{(1)} \cup \cdots \cup \tilde{\mathbf{V}}^{(P+1)}$, with

$$\begin{array}{lllll} \tilde{\mathbf{V}}^{(1)} & = \{V_1^{(1)}, & V_2^{(1)}, & \cdots, & V_m^{(1)}\}, \\ \tilde{\mathbf{V}}^{(2)} & = \{V_1^{(2)}, & V_2^{(2)}, & \cdots, & V_m^{(2)}\}, \\ \cdots & \cdots & & & \\ \tilde{\mathbf{V}}^{(P+1)} & = \{V_1^{(P+1)}, & V_2^{(P+1)}, & \cdots, & V_m^{(P+1)}\}, \end{array}$$

where $V_i^{(k)} = \big(V_i(k), V_i(k+1), \cdots, V_i(T-P+k-1)\big)$, indicating the ith process from the kth time point to the $(T-P+k-1)$th time point, for $i = 1, \cdots, m$ and $k = 1, \cdots, P+1$. After reorganization, in total there are $m * (P+1)$ variables. We constrain that the future cannot cause the past; i.e., variables in $\tilde{\mathbf{V}}^{(i)}$ cannot cause variables in $\tilde{\mathbf{V}}^{(j)}$ when $i > j$.

Figure 6 gives an illustration of a two-variable case with time lag $P = 1$. Figure 6(a) gives the repetitive causal graph over two processes $\mathbf{V}(t) = (V_1(t), V_2(t))^{\mathrm{T}}$ with $t = 1, \cdots, T$. Figure 6(b) gives the unit causal graph over the reorganized set of variables $\tilde{\mathbf{V}} = \tilde{\mathbf{V}}^{(1)} \cup \tilde{\mathbf{V}}^{(2)}$. In this case, we aim to recover the (time-varying) instantaneous causal relations between $V_1^{(2)}$ and $V_2^{(2)}$, and the (time-varying) lagged causal relations from $V_1^{(1)}$ to $V_2^{(2)}$, $V_1^{(1)}$ to $V_1^{(2)}$, and $V_2^{(1)}$ to $V_2^{(2)}$. Note that in this example, when inferring the instantaneous causal relation between $V_1^{(2)}$ and $V_2^{(2)}$, we should consider the influence of the lagged common cause $V_1^{(1)}$.

We also add the surrogate variable C into the causal system to characterize distribution shifts. Algorithm 5 extends Algorithm 1 to recover both instantaneous and lagged causal skeletons and detect changing causal modules. The main procedure is as follows. We first construct a complete undirected graph over the reconstructed variable set $\tilde{\mathbf{V}}$ and C. Then in step 2, we detect changing modules for variables in $\tilde{\mathbf{V}}^{(P+1)}$ by testing the independence between $V_i^{(P+1)}$ and C given a subset of $\tilde{\mathbf{V}}^{(P+1)} \backslash V_i^{(P+1)}$, for $i = 1, \cdots, m$. If they are independent, we remove the edge between $V_i^{(k)}$ and C, for $k = 1, \cdots, P+1$. In step 3, we recover lagged causal relations between variables in $\tilde{\mathbf{V}}^{(P+1)}$ and those in $\tilde{\mathbf{V}}^{(P-p+1)}$, for $p = 1, \cdots, P$. For example, if $V_i^{(P+1)}$ and $V_j^{(P-p+1)}$ are independent given a subset of $\tilde{\mathbf{V}} \backslash \{V_i^{(P+1)}, V_j^{(P-p+1)}\} \cup C$, then remove the edge between $V_i^{(P+1-k)}$ and $V_j^{(P-p+1-k)}$ for $k = 0, \cdots, P-p$. In step 4, we estimate the instantaneous causal skeleton between $V_i^{(P+1)}$ and $V_j^{(P+1)}$ $(i \neq j)$. If $V_i^{(P+1)}$ and $V_j^{(P+1)}$ are independent given a subset of



(a) Repetitive causal graph. (b) Unit causal graph.

Figure 6: A two-variable case with both instantaneous and one-lagged ($P = 1$) causal relationships. (a) Repetitive causal graph. (b) Unit causal graph.

$\tilde{\mathbf{V}} \backslash \{V_i^{(P+1)}, V_j^{(P+1)}\} \cup C \cup \mathbf{Z}_{ij}$, where $\mathbf{Z}_{ij} \subseteq \{\tilde{\mathbf{V}}^{(k)}\}_{k=1}^{P}$ are the lagged common causes of $V_i^{(P+1)}$ and $V_j^{(P+1)}$, then we remove the edge between $V_i^{(k)}$ and $V_j^{(k)}$ for $k = 1, \cdots, P+1$. Note that it is important to consider lagged common causes $\mathbf{Z}_{ij}$ in this step, e.g., $V_1^{(1)}$ is the lagged common cause of $V_1^{(2)}$ and $V_2^{(2)}$ in Figure 6 (b).

For the recovery of causal directions, lagged causal relations obey the rule that past causes future, which is obvious, while instantaneous causal directions can be inferred in the same way as that described in Algorithm 2 and 3.

6.2 With Stationary Confounders

In this section, we discuss the case when there exist stationary confounders; that is, the distribution of the confounder is fixed. We find that distribution shifts over observed variables may still help estimate causal directions in such a case.

To show how distribution shifts help, we analyze two-variable cases in Figure 7, where V_1 and V_2 are observed variables, and Z is a hidden variable which influences both V_1 and V_2. In Figure 7(a) & (b), the causal module of V_1 changes, with changing parameter $\theta_1(C)$, while causal modules of V_2 and Z are fixed. In such a case, we have $V_1 \not\perp\!\!\!\perp V_2$ and $V_1 \not\perp\!\!\!\perp V_2|C$ in both graphs. Furthermore, in (a) we have $C \not\perp\!\!\!\perp V_2$ and $C \not\perp\!\!\!\perp V_2|V_1$, while in (b) we have $C \perp\!\!\!\perp V_2$ and $C \not\perp\!\!\!\perp V_2|V_1$. We can see that distribution shift helps to distinguish between the two graphs when only one causal module changes.

In Figure 7(c) & (d), both the causal modules of V_1 and V_2 change, and they change independently, while the distribution of Z is fixed. In these two graphs, we have $V_1 \not\perp\!\!\!\perp V_2$ and $V_1 \not\perp\!\!\!\perp V_2|C$, and we have $P(V_1) \not\perp\!\!\!\perp P(V_2|V_1)$ and $P(V_2) \not\perp\!\!\!\perp P(V_1|V_2)$, for the reason given below. Hence, graphs in (c) and (d) have the same independence over $\mathbf{V} \cup C$ and between distribution modules, and thus we cannot distinguish between (c) and (d) without further information.

The reason that $P(V_1) \not\perp\!\!\!\perp P(V_2|V_1)$ in (c) is as follows. In (c), $P(V_2|V_1)$ can be represented as: $P(V_2|V_1) = \int P(V_2|V_1, Z)P(Z|V_1)\, dZ$, where $P(V_2|V_1, Z)$ contains the information of $\theta_2(C)$, and $P(Z|V_1)$ contains the information of $\theta_1(C)$, so $P(V_2|V_1)$ also contains the information of $\theta_1(C)$. Since both $P(V_2|V_1)$ and $P(V_1)$ contain the information of $\theta_1(C)$, generally speaking, $P(V_2|V_1)$ and $P(V_1)$ are not independent. Therefore, although, according to the causal module independence, $P(V_1|Z) \perp\!\!\!\perp P(V_2|V_1, Z)$, $P(V_1)$ and $P(V_2|V_1)$ are dependent. Similarly, we can derive that $P(V_2) \not\perp\!\!\!\perp P(V_1|V_2)$ in (d).

Algorithm 5 Detection of Changing Modules and Recovery of Causal Skeleton with both (Time-Varying) Instantaneous and Lagged Causal Relationships

1. Build a complete undirected graph $U_{\mathcal{G}}$ over the variable set $\tilde{\mathbf{V}} \cup C$.
2. (*Detection of changing modules*) For each i ($i = 1, \cdots, m$), test for the marginal and conditional independence between $V_i^{(P+1)}$ and C. If they are independent given a subset of $\tilde{\mathbf{V}}^{(P+1)} \backslash V_i^{(P+1)}$, remove the edge between $V_i^{(P+1)}$ and C in $U_{\mathcal{G}}$; at the same time, remove the edge between $V_i^{(k)}$ ($k = 1, \cdots, P$) and C.
3. (*Recovery of lagged causal skeleton*) For the pth ($p = 1, \cdots, P$) lagged causal relationships, test for the marginal and conditional independence between the variables in $\tilde{\mathbf{V}}^{(P+1)}$ and those in $\tilde{\mathbf{V}}^{(P-p+1)}$. Particularly, for $V_i^{(P+1)}$ and $V_j^{(P-p+1)}$, if they are independent given a subset of $\tilde{\mathbf{V}} \backslash \{V_i^{(P+1)}, V_j^{(P-p+1)}\} \cup C$, remove the edge between $V_i^{(P+1)}$ and $V_j^{(P-p+1)}$; at the same time, remove the edge between $V_i^{(P-k+1)}$ and $V_j^{(P-p+1-k)}$, for $k = 0, \cdots, P-p$.
4. (*Recovery of instantaneous causal skeleton*) For every $i \neq j$, test for the marginal and conditional independence between $V_i^{(P+1)}$ and $V_j^{(P+1)}$. Let $\mathbf{Z}_{ij} \subseteq \{\tilde{\mathbf{V}}^{(k)}\}_{k=1}^{P}$ be the lagged common causes of $V_i^{(P+1)}$ and $V_j^{(P+1)}$, which can be derived from the output of Step 3 above. If $V_i^{(P+1)}$ and $V_j^{(P+1)}$ are independent given a subset of $\tilde{\mathbf{V}}^{(P+1)} \backslash \{V_i^{(P+1)}, V_j^{(P+1)}\} \cup C \cup \mathbf{Z}_{ij}$, remove the edge between $V_i^{(P+1)}$ and $V_j^{(P+1)}$ in $U_{\mathcal{G}}$; at the same time, remove the edge between $V_i^{(k)}$ and $V_j^{(k)}$, for $k = 1, \cdots, P$.

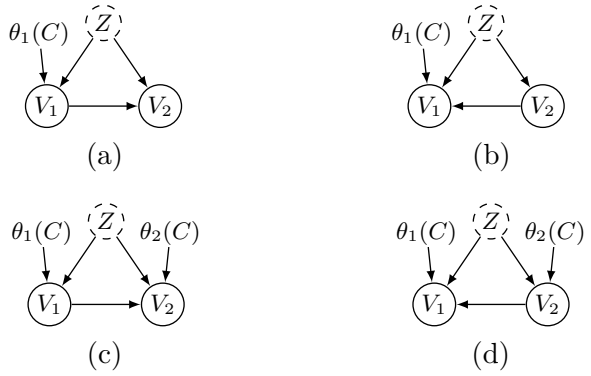


Figure 7: Two-variable cases with stationary confounders. (a) $V_1 \to V_2$ with the changing causal module of V_1. (b) $V_1 \leftarrow V_2$ with the changing causal module of V_1. (c) $V_1 \to V_2$ with the changing causal modules of both V_1 and V_2. (d) $V_1 \leftarrow V_2$ with the changing causal modules of both V_1 and V_2.

6.3 Combination with Approaches Based on Constrained Functional Causal Models

For a pair of adjacent variables V_i and V_j, if none of the identifiability conditions given in Theorem 2 are satisfied, the causal direction between V_i and V_j is undetermined with CD-NOD. In such a case, to infer the causal direction between V_i and V_j, one possible way is to further leverage the approaches based on constrained functional causal models, e.g., the linear non-Gaussian model (Shimizu et al., 2006), the nonlinear additive noise model (Hoyer et al., 2009; Zhang and Hyvärinen, 2009a), or the post-nonlinear model (Zhang and Chan, 2006; Zhang and Hyvärinen, 2009b). To determine the causal direction with functional causal model-based approaches, we have two scenarios to consider, depending on whether $P(V_i, V_j)$ changes or not.

1. The joint distribution $P(V_i, V_j)$ is fixed. In this scenario, we can fit a constrained functional causal model, e.g., the nonlinear additive noise model, as usual. Let $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_i, V_j\}$ be a minimal deconfounding set of V_i and V_j. We

   - first assume $V_i \to V_j$ and fit an additive noise model

     $$V_j = f_j(V_i, \mathbf{Z}) + E_j,$$

     and test the independence between estimated residual $\hat{E}_j$ and hypothetical causes $(V_i, \mathbf{Z})$,
   - and then assume $V_j \to V_i$ and fit an additive noise model in this direction

     $$V_i = f_i(V_j, \mathbf{Z}) + E_i,$$

     and test the independence between estimated residual $\hat{E}_i$ and hypothetical causes $(V_j, \mathbf{Z})$.

   We choose the direction in which the independence between estimated residual and hypothetical causes holds.

2. The joint distribution $P(V_i, V_j)$ changes. The fact that for V_i and V_j, none of the conditions in Theorem 2 is satisfied implies that neither V_i nor V_j is adjacent to C. It further infers that there are some variables in $\mathbf{V}$ with changing distributions and influencing V_i or V_j. Let $\mathbf{Z}_i \subseteq \mathbf{V} \backslash \{V_i, V_j\}$ represent a set of variables which directly influence V_i, and satisfies that $\forall Z_i \in \mathbf{Z}_i$, its distribution $P(Z_i)$ changes. Similarly, we denote $\mathbf{Z}_j$ as parents of V_j that have changing distributions. Let $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_i, V_j\}$ be a minimal deconfounding set of V_i and V_j, and let $\tilde{\mathbf{Z}} \subseteq \mathbf{Z}$, with $P(\tilde{\mathbf{Z}})$ fixed and $\forall Z \in \mathbf{Z} \backslash \tilde{\mathbf{Z}}$, $P(Z)$ changes. Note that it is easy to infer that $\mathbf{Z}_i \subseteq \mathbf{Z}$ and $\mathbf{Z}_j \subseteq \mathbf{Z}$. We

   - first fit an additive noise model by assuming $V_i \to V_j$

     $$V_j = f_j(V_i, \tilde{\mathbf{Z}}, \mathbf{Z}_j) + E_j,$$

     and test the independence between estimated residual $\hat{E}_j$ and hypothetical causes $(V_i, \mathbf{Z}_j, \tilde{\mathbf{Z}})$,

– and then assume $V_j \to V_i$ and fit an additive noise model accordingly

$$V_i = f_i(V_j, \tilde{\mathbf{Z}}, \mathbf{Z}_i) + E_i,$$

and test the independence between estimated residual $\hat{E}_i$ and hypothetical causes $(V_j, \mathbf{Z}_i, \tilde{\mathbf{Z}})$.

We choose the direction in which the independence between estimated residual and hypothetical causes holds. Note that in case 2, we cannot ignore the changing factors $\mathbf{Z}_i$ and $\mathbf{Z}_j$ in their corresponding functional causal models, because functional causal model-based approaches assume a fixed causal model. If we do not consider $\mathbf{Z}_i$ or $\mathbf{Z}_j$, then the corresponding functions f_i and f_j, respectively, may change across data sets or over time.

If there are unoriented edges to V_i or V_j, the potential minimal deconfounding set $\tilde{\mathbf{Z}}$, and potential changing influences $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are chosen in a heuristic way, similar to that in Algorithm 3.

7. Relations between Heterogeneity/Nonstationarity and Soft Intervention in Causal Discovery

In this section, we show that heterogeneity/nonstationarity and soft intervention are tightly related in causal discovery, and the former may be more effective for causal discovery than the latter. The definition of soft intervention is as follows.

Definition 4 (Soft Intervention (Eberhardt and Scheines, 2007)) *Given a set of measured variables $\mathbf{V}$, a soft intervention I_p on a variable $S \in \mathbf{V}$ is an intervention on S that satisfies the following constraints: when $I_p = 1$, I_p does not make S independent of their causes in $\mathbf{V}$; i.e., it does not break any edges that are incident to S. The causal module $P(S|\mathrm{PA}^S)$ is replaced by $P^*(S|\mathrm{PA}^S, I_p = 1)$, where*

$$P^*(S|\mathrm{PA}^S, I_p = 1) \neq P(S|\mathrm{PA}^S, I_p = 0). \tag{19}$$

Otherwise, all terms remain unchanged.

Additionally, I_p satisfies the following properties:

- *I_p is a variable with two states, $I_p = 1$ or $I_p = 0$, whose state is known.*
- *When $I_p = 0$, the passive observational distribution over S obtains.*
- *I_p is a direct cause of S and only S.*
- *I_p is exogenous, that is, uncaused.*

Note that the soft intervention does not imply any structural changes over the variables in $\mathbf{V}$; instead, it influences the manipulated probability distribution.

Heterogeneity/nonstationarity can be seen as a consequence of soft interventions done by nature. Recall that the heterogeneous/nonstationary causal modules are defined as

$$V_i = f_i\big(\mathrm{PA}^i, \mathbf{g}^i(C), \theta_i(C), \epsilon_i\big),$$

where $\mathbf{g}^i(C) \subseteq \{g_l(C)\}_{l=1}^L$ denotes the set of confounders that influence V_i (it is an empty set if there is no confounder behind V_i and any other variable), $\theta_i(C)$ denotes the effective parameters in the model that are also assumed to be functions of C, and ϵ_i is a disturbance term that is independent of C. The distribution of $P(V_i|\mathrm{PA}^i, C)$ changes when C takes different values, i.e.,

$$P(V_i|\mathrm{PA}^i, C = c) \neq P(V_i|\mathrm{PA}^i, C = c'). \tag{20}$$

This means that across the values of C, c and c', there exists a soft intervention on V_i.

Our developed framework in causal discovery from heterogeneous/nonstationary data is more applicable than that from Eberhardt and Scheines (2007) in the following aspects:

1. In Eberhardt and Scheines (2007), I_p is known and only contains two states, while in our case we can detect where it happens in an automated way, and the changes can be discrete (across domains) or continuous (over time).
2. In Eberhardt and Scheines (2007), I_p is assumed to be a direct cause to a single variable, while in our case there is no such a restriction. For example, we allow pseudo confounders $g_l(C)$ which affects several variables.
3. Our framework also allows edges to vanish or appear in some domains or over some time periods; that is, the structure over $\mathbf{V}$ may change, for which hard intervention is a special case, while in Eberhardt and Scheines (2007), the causal structure does not change.

8. Experimental Results

We applied the proposed CD-NOD (both phase I and phase II) to both synthetic and real-world data sets to learn causal graphs and identify changing causal modules, and then we learned a low-dimensional representation of changing causal modules (phase III).

8.1 Synthetic Data

8.1.1 Setting 1

To show the generality of the proposed methods in causal discovery when data distributions change, we considered two types of data: multi-domain heterogeneous data and nonstationary time series. Moreover, we considered changes in both causal strength and noise variances.

For variable V_i whose causal module changes, the tth data point was generated according to the following functional causal model:

$$V_{i,t} = \sum_{V_j \in \mathrm{PA}^i} b_{ij,t} f_i(V_{j,t}) + \sigma_{i,t}\epsilon_{i,t},$$

where PA^i denotes the set of V_i's direct causes, $V_j \in \mathrm{PA}^i$ is the jth direct cause of V_i, $b_{ij,t}$ is the varying causal strength from variable V_j to V_i, and $\sigma_{i,t}$ is the changing parameter applied to the noise term $\epsilon_{i,t}$. The function f_i was randomly chosen from linear, cubic, tanh, sinc functions, or random mixtures of those functions. The noise term $\epsilon_{i,t}$ was randomly chosen from a uniform distribution $\mathcal{U}(-0.5, 0.5)$ or a standard normal distribution $\mathcal{N}(0, 1)$.

The varying causal strength $b_{ij,t}$ and varying noise's parameter $\sigma_{i,t}$ were generated in different ways for heterogeneous and nonstationary data.

- For heterogeneous data whose data distributions change across domains, in each domain, $b_{ij,t}$ and $\sigma_{i,t}^2$ were randomly generated from uniform distributions $\mathcal{U}(0.5, 2.5)$ and $\mathcal{U}(1, 3)$, respectively. Note that in the same domain both $b_{ij,t}$ and $\sigma_{i,t}^2$ remain the same.

- For nonstationary data whose data distributions smoothly change over time, we generated $b_{ij,t}$ and $\sigma_{i,t}$ by sampling from a Gaussian process (GP) prior $b_{ij}(\mathbf{t}) \sim \mathcal{GP}(\vec{0}, K(\mathbf{t}, \mathbf{t}))$ and $\sigma_i(\mathbf{t}) \sim \mathcal{GP}(\vec{0}, K(\mathbf{t}, \mathbf{t}))$, respectively. We used squared exponential kernel $k(t, t') = \exp(-\frac{1}{2}\frac{|t-t'|}{\lambda^2})$, where λ is the kernel width.

For variable V_i whose causal module is fixed, it was generated according to a fixed functional causal model:

$$V_{i,t} = \sum_{V_j \in \mathrm{PA}^i} b_{ij} f_i(V_{j,t}) + \epsilon_{i,t},$$

where both the causal strength b_{ij} and variance of the noise term ϵ_i are fixed over time.

We randomly generated acyclic causal structures according to the Erdos-Renyi model (Erdős and Rényi, 1959) with the probability of each edge 0.3. Each generated graph has 6 variables. In each graph, we randomly picked variables with changing causal modules, and the number was randomly chosen from $\{4, 5, 6\}$. We also considered different sample sizes. Particularly, for heterogeneous data, the sample size of each domain was chosen between 50 and 100, and the total sample size was N =600, 900, 1200, and 1500. For nonstationary data, we fixed the kernel with of the Gaussian prior and generated the data with sample size $N = 600, 900, 1200,$ and 1500. In each scenario (each data type and each sample size), we generated 100 random realizations.

We applied the proposed CD-NOD to identify the underlying causal structure, both causal skeleton (phase I) and causal directions (phase II), and detect changing causal modules (phase I).

Phase I: causal skeleton identification and changing causal module detection. We applied CD-NOD phase I (Algorithm 1) to identify the causal skeleton and detect changing causal modules. Specifically, for heterogeneous data, we used the domain index as the surrogate variable C to capture the distribution change, while for nonstationary data, we used the time index. We used PC (Spirtes et al., 1993) as the search procedure and the kernel-based conditional independence (KCI) test to test conditional independence relationships.

We compared our approach with the original constraint-based method, which does not take into account distribution changes. For the original constraint-based method, we applied the PC search over the observed variables, combined with the KCI test. We also compared with approaches based on the minimal change principle, the identical boundaries (IB) method and the minimal changes (MC) method (Ghassami et al., 2018). Both IB and MC methods are designed for multi-domain causal discovery in linear systems.

For the KCI test, the significance level was 0.05. We used Gaussian kernels, and the hyperparameters, e.g. the kernel width, were set with empirical values; please refer to Zhang et al. (2011) for details. Since both IB and MC methods need data from multiple domains, for heterogeneous data, we segmented the data according to the domain index

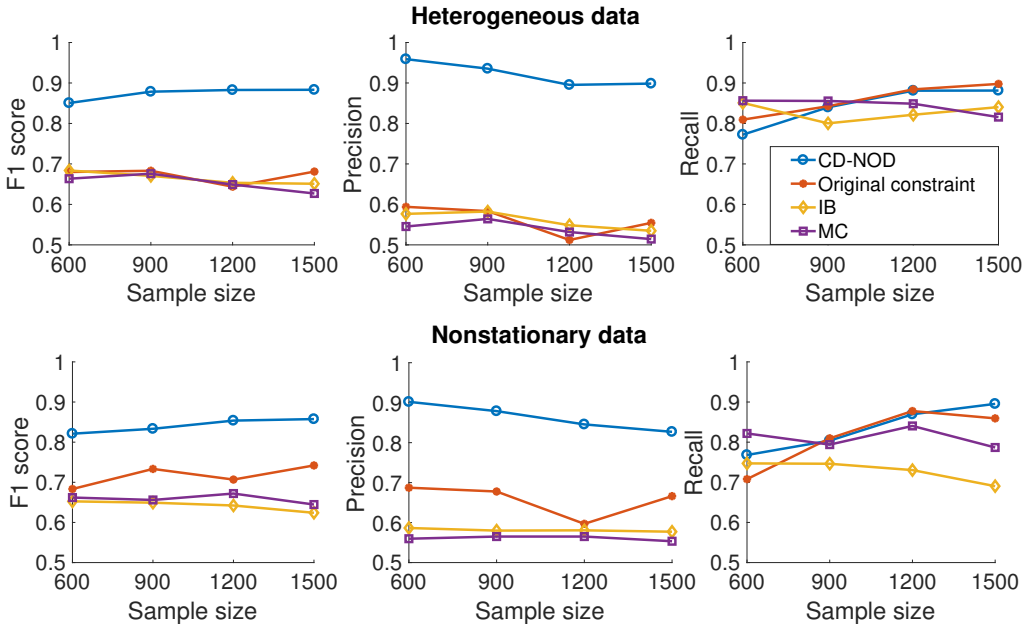


Figure 8: Accuracy of the recovered causal skeleton from heterogeneous (upper row) and nonstationary (lower row) data. We compared three accuracy measures: F_1 score (left column), precision (middle column), and recall (right column). We compared our proposed CD-NOD (phase I) with original constraint-based method, IB method, and MC method.

before applying IB and MC methods, and for nonstationary data, we segmented the data into non-overlapping domains, with sample size 100 in each domain.

Figure 8 gives the accuracy of the recovered causal skeleton from both heterogeneous (upper row) and nonstationary (lower row) data. The x-axis indicates the sample size, and the y-axis shows the accuracy of the recovered causal skeleton. We considered three accuracy measures: F_1 score, precision, and recall, where precision indicates the rate of spurious edges, recall the rate of missing edges, and $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. We can see that the proposed CD-NOD (phase I) gives the best F_1 score and precision on both heterogeneous and nonstationary data for all sample sizes, and it gives similar recall with others. Moreover, there is a slight increase in F_1 score and recall along with the sample size. The original constraint-based method gives a much lower precision, and hence a lower F_1 score as well, which means that there are many spurious edges. The reason is that it does not consider distribution shifts, as we illustrated in Figure 1. Both IB and MC methods give a much lower precision, probably because they are designed for linear systems, and thus the indirect nonlinear influences cannot be totally captured by variables along the pathway with linear functions.

Table 2 shows the accuracy of the detected changing causal modules, measured by F_1 score, precision, and recall. We can see that the proposed CD-NOD performs well, and that the F_1 score tends to increase along with sample size in both scenarios.

N	F_1	Precision	Recall
600	0.93	0.98	0.89
900	0.95	0.99	0.93
1200	0.96	1.00	0.93
1500	0.96	0.99	0.93

(a) Heterogeneous data

N	F_1	Precision	Recall
600	0.94	0.99	0.90
900	0.96	0.99	0.94
1200	0.97	0.99	0.96
1500	0.96	0.97	0.95

(b) Nonstationary data

Table 2: Accuracy of the identified changing causal modules from (a) heterogeneous data and (b) nonstationary data.

Phase II: causal direction identification. We then applied three rules to identify causal directions, based on the skeleton learned in phase I. The three rules are

(1) generalization of invariance (Algorithm 2),

(2) independent changes between causal modules (Algorithm 3),

(3) and Meek's orientation rule (Meek, 1995).

The three rules are applied in the above sequence to identify causal directions.

We compared with the window-based method (Zhang et al., 2017), the IB method, and the MC method for direction determination. The window-based method identifies the causal direction by measuring the dependence between hypothetical causal modules; however, the kernel density estimation of modules were performed in each sliding window or in each domain, separately. For the window-based method, we used the causal skeleton derived from CD-NOD phase I. For heterogeneous data, the window size was the same as the sample size in each domain; for nonstationary data, the size of each sliding window was 100. For the IB and MC method, the skeleton and orientations were derived simultaneously.

Figure 9 shows the F_1 score of the recovered whole causal graph (both skeletons and directions) from heterogeneous and nonstationary data. We can see that the proposed CD-NOD gives the best F_1 score in all scenarios. The window-based method has second-best performance, and it performs slightly better on heterogeneous data than on nonstationary data. The reason may be that on nonstationary data, the sliding window-based method may lead to large estimation errors, especially when the causal influence varies quickly over time. The performance of IB and MC methods is worse; it is not surprising since the generated data are not linear.

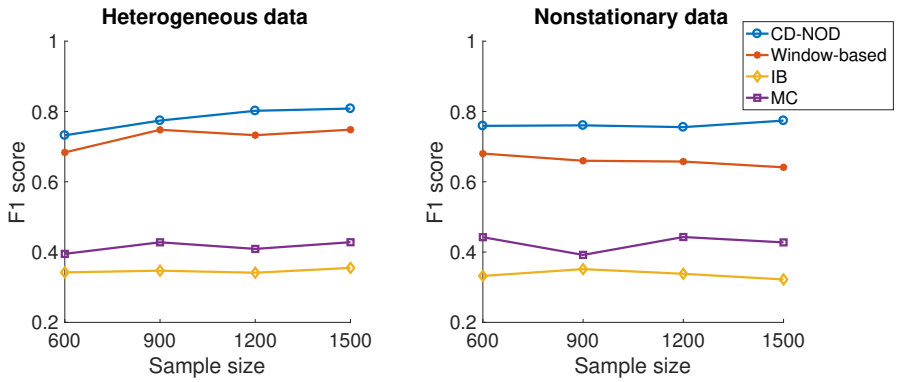


Figure 9: Accuracy of the recovered whole causal graph (both skeletons and directions) from both heterogeneous and nonstationary data. We compared our proposed CD-NOD with window-based method, IB method, and MC method.

8.1.2 Setting 2

To clearly show the efficacy of Algorithm 3 in causal direction identification (without being affected by the accuracy from causal skeleton identification and the other two orientation rules), we generated another synthetic data set. Particularly, we generated fully connected acyclic graphs. For each variable V_i, all its causal strength b_{ij} and noise's parameter σ_i change, either across domains or over time. We considered different graph sizes $m = 2, 4,$ and 6 and different sample sizes $N = 600, 900, 1200,$ and 1500.

To assess the performance of causal direction identification, we assumed that the causal skeleton is given (fully connected), and we inferred causal directions with Algorithm 3, which exploits the independence between causal modules. We compared CD-NOD with the window-based method, the IB method, and the MC method. The setting of hyperparameters and window size was the same as that in Setting 1.

Figure 10 gives the accuracy (F_1 score) of the inferred causal directions from heterogeneous and nonstationary data. We can see that CD-NOD (Algorithm 3) gives the best F_1 score, and its accuracy slightly increase with sample size. The accuracy of window-based method is comparably lower than CD-NOD (Algorithm 3), and its performance is better on heterogeneous data than that on nonstationary data. The accuracy of the IB and MC method is much lower, since they are only for linear systems.

8.1.3 Setting 3

We further investigated the effect of sample size per domain in heterogeneous data of the final performance. We varied the number of samples N_0 in each domain, with $N_0 = 10, 20, 40, 60, 80,$ and 100, and there were in total 10 domains. Other settings of the data generating process are the same as that in Setting 1.

Figure 11(left) gives the F_1 score of the recovered causal skeleton, along with the number of samples per domain. We compared CD-NOD with original constraint-based methods

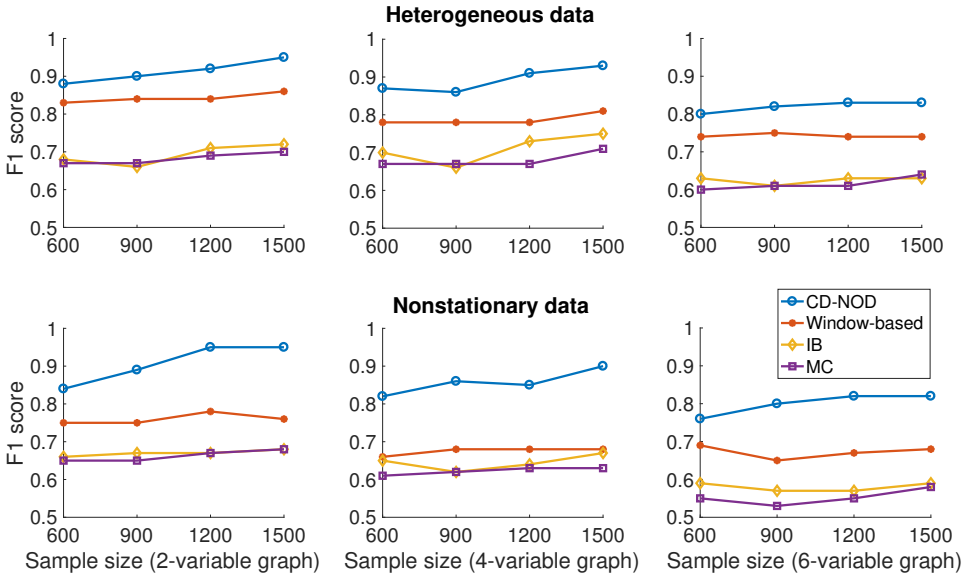


Figure 10: Accuracy of inferred causal directions from heterogeneous (upper row) and nonstationary (lower row) data. We tested the accuracy on 2-variable (left column), 4-variable (middle column), 6-variable (right column) fully connected graphs. We compared our proposed CD-NOD with the window-based method, the IB method, and the MC method.

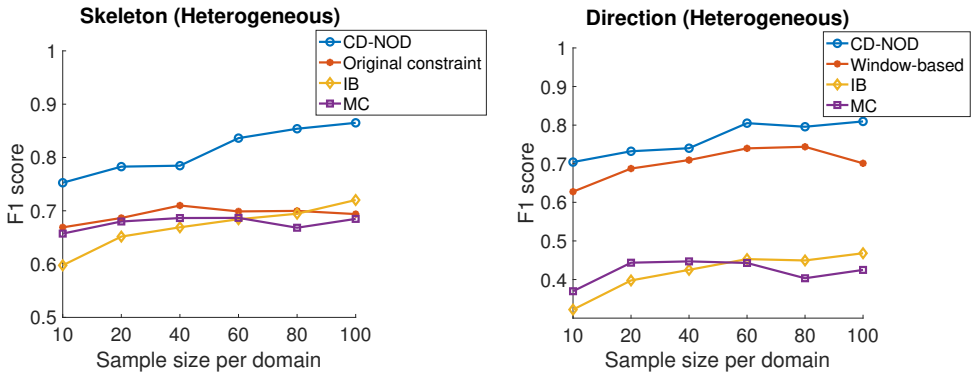


Figure 11: Accuracy of the estimated causal skeleton (left) and the whole causal graph (right) from heterogeneous data, along with the number of samples in each domain.

(without taking into account the distribution shift), the IB method, and the MC method. Figure 11(right) shows the F_1 score of the recovered whole causal graph, including both the skeleton and directions, produced by CD-NOD, the window-based method, the IB method, and the MC method, respectively. We can see that CD-NOD achieves the best accuracy in all cases, regarding both the estimated skeleton and the whole causal graph. Moreover, we found that not surprisingly, its performance improves with the number of samples in each domain, with a relatively large improvement (around 5%) from $N_0 = 40$ to $N_0 = 60$.

8.1.4 Setting 4

After identifying the causal structure, we then estimated the nonstationary driving force of changing causal modules by the procedure KNV given in Algorithm 4. For illustration purposes, we used a single driving force for each module, which changes both causal strength b and noise parameter σ.

Phase III: nonstationary driving force estimation We used Gaussian kernels both in kernel embedding of constructed joint distributions and kernel PCA. We compared our approach with the linear time-dependent functional causal model (Huang et al., 2015), which puts a GP prior on time-varying coefficients and uses the mean of the posterior to represent the nonstationary driving force. In addition, we compared our methods with Bayesian change point detection (Adams and MacKay, 2007), [4] which is widely used in nonstationary data to detect change points.

Figure 12(a) and 12(b) show the estimated low-dimensional driving force of changing causal modules for smooth changes (nonstationary data) and sudden changes (heterogeneous data), respectively, when $N = 600$. In left panels, blue lines show the estimated driving force by KNV, and red lines show ground truth. Vertical black dashed lines indicate detected change points by Bayesian change point detection. Middle panels show the largest ten

4. We used the implemented Matlab code from http://hips.seas.harvard.edu/content/bayesian-online-changepoint-detection.

eigenvalues of Gram matrix M^g. In right panels, blue lines are the recovered changing components by the linear time-dependent GP, and red lines are ground truth. The scale of the y-axis has been adapted for clarity. We only showed the first principal component in the left panel, since the first eigenvector captures most of the variance, as shown in middle panels. We can see that KNV gives the best recovery of the changing components in all cases. Bayesian change point detection, which aims to estimate change points in the marginal or joint distributions (instead of causal mechanisms), fails to handle the case of smooth changes. The linear time-dependent GP does not work well, and one reason is that it cannot account for influences from changing noise parameters, in cases of both smooth and sudden change.

8.2 Real-World Data Sets

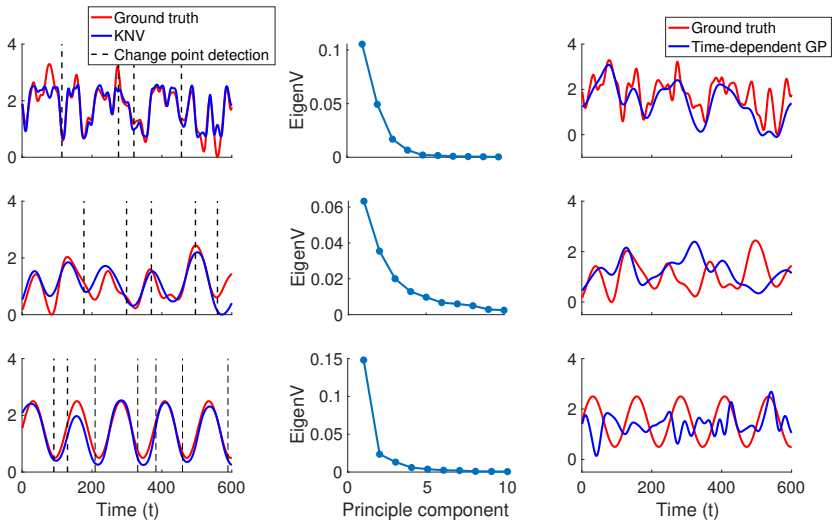
We then applied the proposed CD-NOD to three real-world data sets. We performed the same three procedures as before: recover causal structure over variables, identify changing causal modules, and estimate low-dimensional driving forces of changing causal modules. Specifically, for causal skeleton determination and changing module detection, we applied Algorithm 1. For causal direction determination, we applied three orientation rules: Algorithm 2, Algorithm 3, and Meek's orientation rule, in sequence. For driving force estimation, we applied KNV (Algorithm 4). Since our proposed methods outperform all the alternatives on synthetic data, we only applied our proposed methods to real-world data sets.

8.2.1 Task fMRI

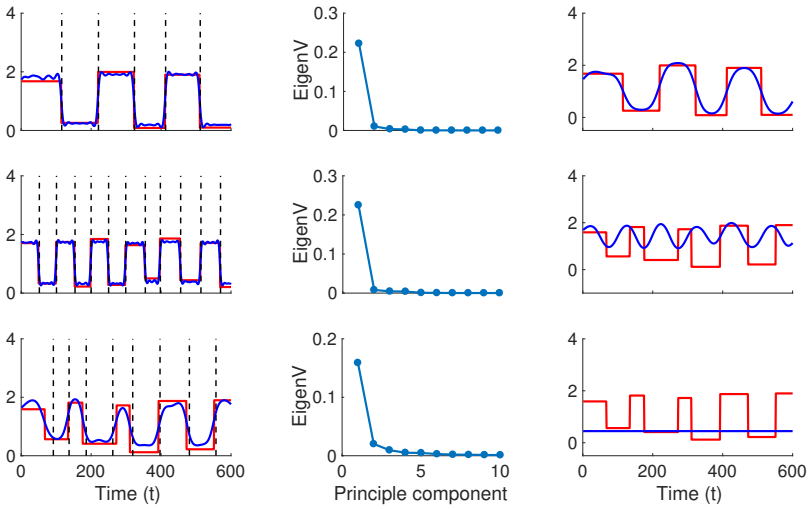
We applied our methods to task fMRI data, which were recorded under a star/plus experiment (Wang et al., 2004). There are three states in the experiment. (1) The subject rested or gazed at a fixation point on the screen (State 1). (2) The subject was shown a picture and a sentence that was not negated, and was instructed to press a button to indicate whether the sentence matched the picture (State 2). (3) The subject was shown a picture and a negated sentence, and was instructed to press a button indicating whether the sentence matched the picture (State 3). The time resolution of the recording is 500 ms. The fMRI voxel data are organized into regions of interests (ROIs), resulting in 25 ROIs.

Figure 13 shows the recovered causal structure over the 25 ROIs, where red circles mean that corresponding brain areas have changing causal mechanisms. We can see that the causal edges from the ROIs in the right hemisphere to the corresponding left ones are robust, e.g., RIPL $\rightarrow$ LIPL, RIT $\rightarrow$ LIT. ROIs CALC, LSGA, LIPL, and LIT have large indegree centrality, where CALC and LIT are responsible for visual input processing, and LSGA and LIPL are for language processing. The causal modules of CALC, RIPL, LIPL, RSGA, LSGA, RDLPFC, LDLPFC, RTRIA, LSP, and RIT are time-varying, which are marked with red circles. The identified changing causal modules correspond to key areas for visual and language perception. By considering nonstationarity, 16 connections that were produced by the original PC algorithm are removed, e.g., LIPL - RSGA, LSPL - RDLPFC, and RDLPFC - RTRIA, indicating possible pseudo confounders behind those pairs of ROIs.

Figure 14 visualizes the estimated nonstationary driving force by KNV. The top figure shows the state of input stimuli: State = 1 (State 1) means that the subject was in resting state; State = 2 (State 2) means that the subject was shown a picture and a non-negated sentence, and State = 3 (State 3) means that the shown sentence was negated. The lower



(a) Smooth changes



(b) Sudden changes

Figure 12: Visualization of estimated driving forces of changing causal modules for (a) smooth changes and (b) sudden changes. Left panel: blue lines show the recovered nonstationary components by KNV. Red lines are ground truth. Vertical black dashed lines indicate detected change points by Bayesian change point detection. Middle Panel: the largest ten eigenvalues of Gram matrix M^g. Right Panel: blue lines are recovered nonstationary components by linear time-dependent GP. Red lines are ground truth.

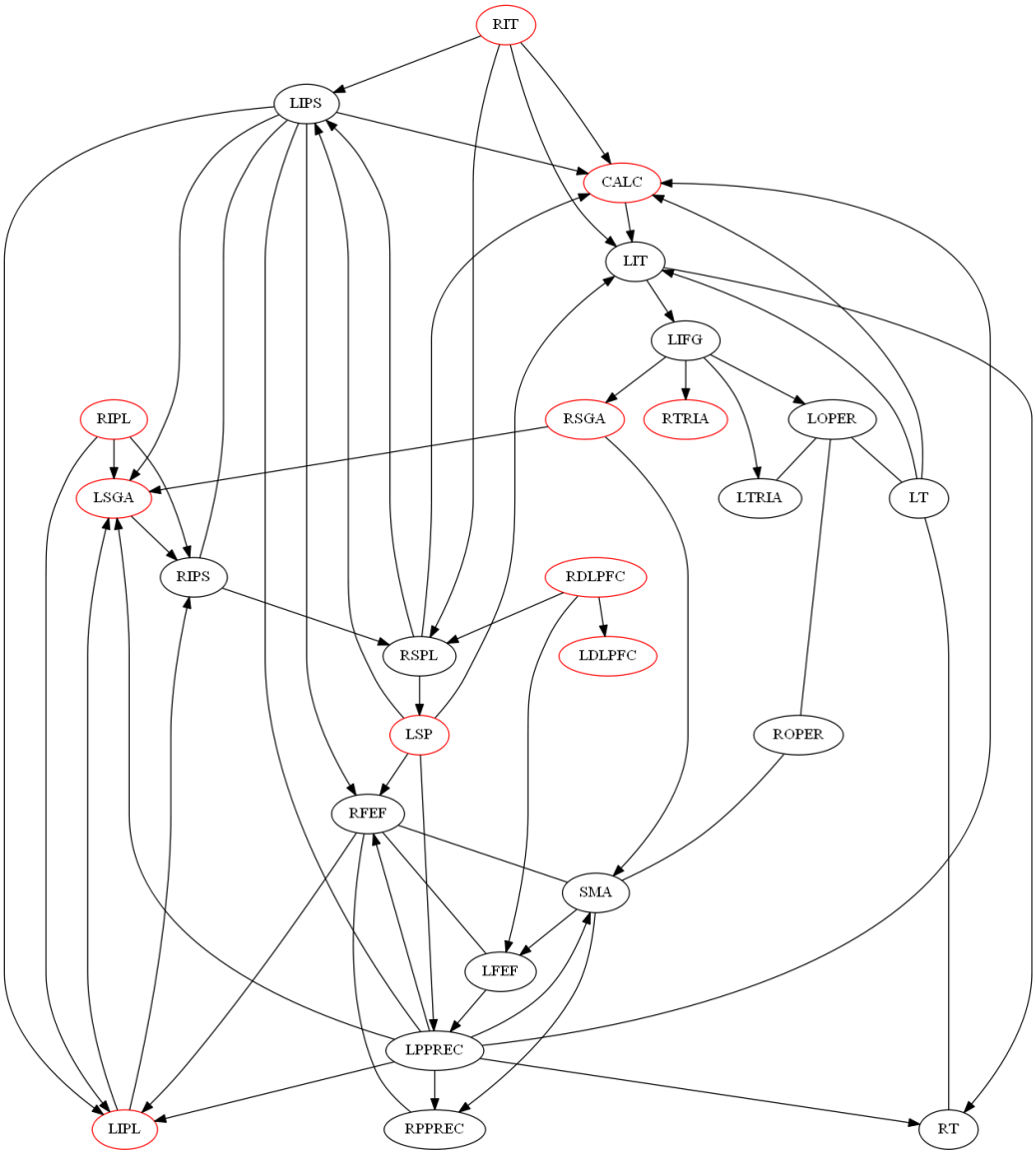


Figure 13: Recovered causal graph over 25 ROIs. The red circles denote that causal modules of corresponding brain areas changing over states.

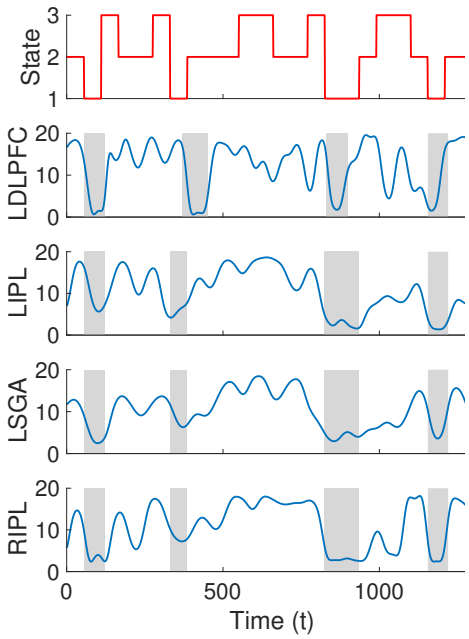


Figure 14: Visualization of estimated nonstationary driving force from task fMRI data. The top figure shows the state of input stimuli: State = 1 means that the subject was in resting state; State = 2 means that the subject was shown a picture and a sentence, in which the sentence was not negated, while in State = 3 the sentence was negated. The other four figures show the estimated nonstationary driving forces of LDLPFC, LIPL, LSGA, and RIPL, respectively. The shaded areas are intervals which have obvious changes; they match the resting states (State = 1) with no input stimuli. For LDLPFC, the second shaded interval is delayed. There are changes between State 2 and State 3 but are less obvious, comparatively.

four figures show the estimated nonstationary driving forces of the causal modules of several key ROIs: LDLPFC, LIPL, LSGA, and RIPL, respectively. The shaded areas are intervals which have obvious changes; they match the resting state (State 1) with no input stimuli. Quantitatively, the correlations between states and the recovered driving forces of LDLPFC, LIPL, LSGA, and RIPL are 0.22, 0.49, 0.41, 0.47, respectively. For LDLPFC, the second shaded interval is delayed, which may be due to the fact that LDLPFC is involved in cognitive processes, e.g., working memory, so that it may remain activated for some time when the input stimulus has been removed. There are changes between State 2 and State 3, but they are less obvious. In short, the recovered nonstationary driving forces show that there are obvious changes between the resting state and the task states, while the changes between two task states are relatively less obvious.

8.2.2 Stock Returns

We applied our methods to daily returns of stocks from Hong Kong (HK) and the United States (US), downloaded from Yahoo Finance.

HK Stock Market The HK stock dataset contains 10 major stocks, which are daily dividend adjusted closing prices from 10/09/2006 to 08/09/2010. For the few days when the stock price is not available, a simple linear interpolation is used to estimate the price. Denoting the closing price of the ith stock on day t by $P_{i,t}$, the corresponding return is calculated by $V_{i,t} = \frac{P_{i,t}-P_{i,t-1}}{P_{i,t-1}}$. The 10 stocks are Cheung Kong Holdings (1), Wharf (Holdings) Limited (2), HSBC Holdings plc (3), Hong Kong Electric Holdings Limited (4), Hang Seng Bank Ltd (5), Henderson Land Development Co. Limited (6), Sun Hung Kai Properties Limited (7), Swire Group (8), Cathay Pacific Airways Ltd (9), and Bank of China Hong Kong (Holdings) Ltd (10). Among these stocks, 3, 5, and 10 belong to Hang Seng Finance Sub-index (HSF), 1, 8, and 9 belong to Hang Seng Commerce & Industry Sub-index (HSC), 2, 6, and 7 belong to Hang Seng Properties Sub-index (HSP), and 4 belongs to Hang Seng Utilities Sub-index (HSU).

Figure 15 shows the estimated causal structure, where the causal modules of $2, 3, 4, 5$ and 7 are found to be time-dependent, indicated by red circles. In contrast, the PC algorithm with the KCI test (without taking into account the nonstationarity) gives five more edges, which are $2-3$, $2-5$, $3-6$, $5-7$, and $4-5$; most of these spurious edges (the former four) are in the finance sub-index and the properties sub-index, indicating the existence of pseudo confounders behind these sub-indices. We found that all stock returns that have changing causal mechanisms are in HSF, HSP, and HSU; they might be affected by some unconsidered changing factors, e.g., the change of economic policies. Furthermore, we estimated causal directions by the procedure given in Algorithm 2 and 3, and the inferred directions are consistent with some background knowledge of the market. For instance, the within sub-index causal directions tend to follow the owner-member relationship. Examples include $5 \to 3$ (note that 3 partially owns 5), $9 \to 8$, and $4 \to 1$. Those stocks in HSF are major causes for those in HSC and HSP, and the stocks in HSP and HSU impact those in HSC. These results indicate that overseas markets (e.g., the US market) may influence the local market via large banks, and that prices of stocks in commerce and industry are usually affected by certain companies in finance, properties, and utilities sectors.

Figure 16 (bottom panels) visualizes the estimated driving forces of stocks $2, 3, 4, 5$, and 7 (the scale of y-axis has been adjusted). We can see that the nonstationary components of root causes, 4 and 5, share a similar variability; the change points are around 01/22/2007 (T_1), 07/16/2007 (T_2), 06/30/2008 (T_3), and 02/11/2009 (T_4). The nonstationary components of the causal modules of 2, 3, and 7 have change points around T_2, T_3, and T_4. Stock 5 has additional change points at T_1 (01/22/2007), which is perhaps due to the fact that Hang Seng Bank established its wholly owned subsidiary Hang Seng Bank (China) Limited in 2007. The change points around T_2, T_3, and T_4 match with the critical time points of financial crisis around the year of 2008. The active phase of the crisis, which manifested as a liquidity crisis, could be dated from August, 2007,[5] around T_2. The estimated driving

5. See more information at `https://en.m.wikipedia.org/wiki/Financial_crisis_of_2007-08`.

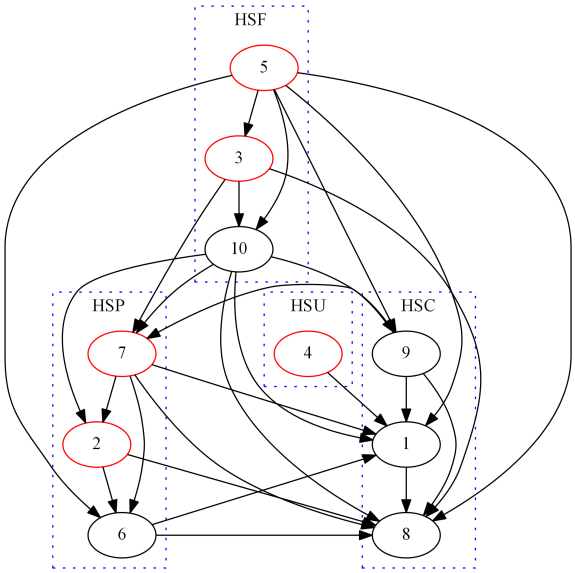


Figure 15: Recovered causal graph over the 10 main HK stocks. Red circles indicate that causal modules of corresponding stocks change over time.

forces are consistent with the change of the TED spread,[6] which is an indicator of perceived credit risk in the general economy and is shown in Figure 16 (top panel) for comparison.

US Stock Market We then applied our methods to daily returns of a number of stocks from New York Stock Exchange, downloaded from Yahoo Finance. We considered 80 major stocks and used their daily dividend adjusted closing prices from 07/05/2006 to 12/16/2009. They are clustered into 10 sectors: energy, public utilities, capital goods, health care, consumer service, finance, transportation, consumer nondurable goods, basic industry, and technology.

Figure 17 shows the estimated causal graph over stock returns, each color representing one sector. The size of each node reflects its degree of connections (sum of indegree and outdegree), and larger node indicates a higher degree. We found that intra-sector connections are more dense than inter-sector connections. The stocks in energy, finance, public utilities, and basic industries are more likely to be causes of stocks in other sectors; among these four sectors, stocks in energy and finance usually causally influence stocks in utilities and basic industries.

More specifically, GE, a stock in energy, does not have causal edges within the sector; instead, it is adjacent to stocks in other sectors, which may be due to the reason that GE is a conglomerate and has various segments, including healthcare, power, transportation, oil and gas, etc. In regard to causal interactions within energy, CHK and KEG are root causes, which may be because KEG is an oilfield services company, and CHK is a holding company for a variety of energy enterprises; NBR and HAL are also more likely to cause others, and PBR, WMB, TLM, and NE are more likely to be influenced by other stocks.

6. See `https://en.m.wikipedia.org/wiki/TED_spread`.

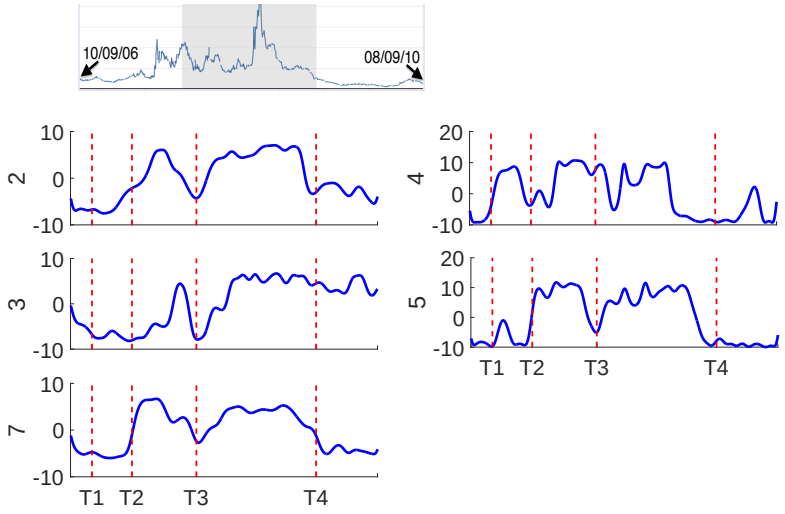


Figure 16: The estimated nonstationary driving forces of time-dependent stock returns, as well as the curve of the TED spread, over the period from 10/09/2006 to 08/09/2010. Top: Curve of the TED spread shown for comparison. Bottom: Estimated nonstationary driving force of stocks $2, 3, 4, 5$, and 7, where T_1, T_2, T_3, and T_4 stand for 01/22/2007, 07/16/2007, 05/03/2008, and 11/02/2009, respectively. We can see that the nonstationary components of root causes, 4 and 5, share the similar variability with change points around T_1, T_2, T_3, and T_4. The nonstationary components of 2, 3, and 7 have change points only around T_2, T_3, and T_4.

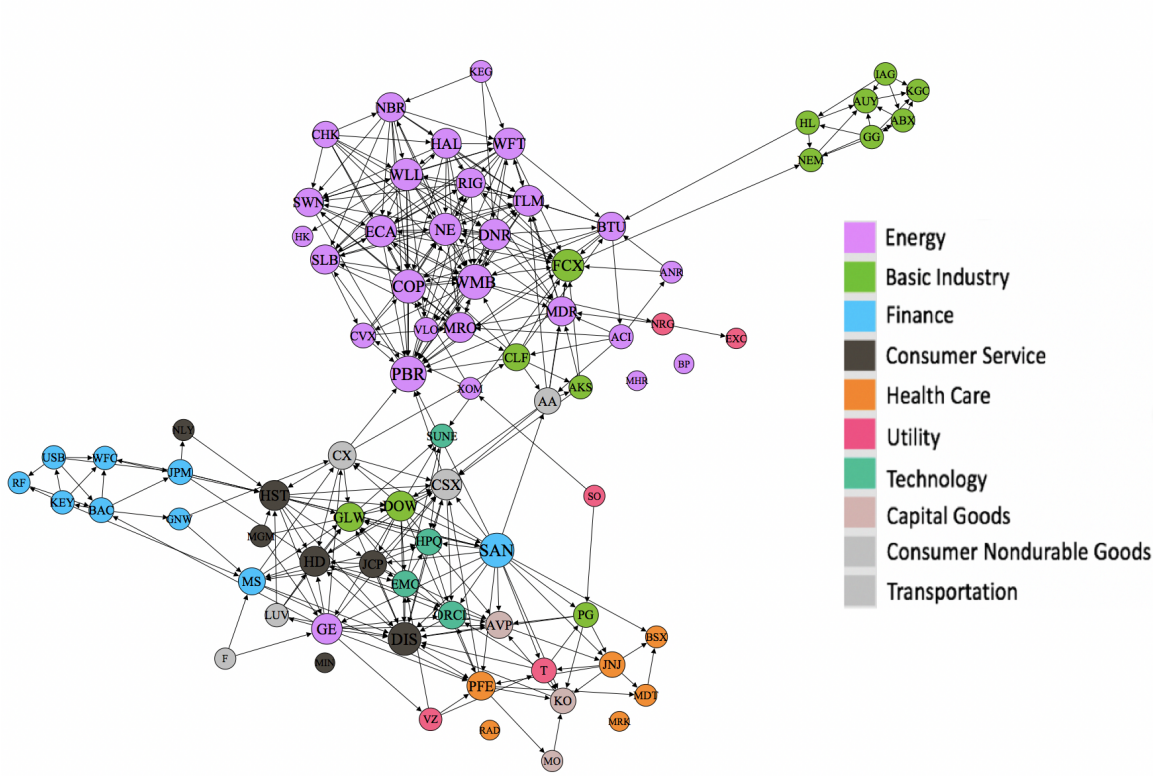


Figure 17: Recovered causal graph over 80 NYSE stocks. Each color of nodes represents one sector.

Among the stocks in finance, MS, SAN, and JPM have a large number of edges with stocks from other sectors. Within finance, USB and KEY usually cause others, and WFC is the sink node. Among the stocks in basic industry, FCX is largely affected by stocks from energy, and it causes other stocks in basic industry, probably because it also belongs to petroleum industry; GLW and DOW have a large number of edges with stocks in consumer service and technology, which may be related to their products in glass & ceramic materials, and they are not adjacent to other stocks in basic industry. Within basic industry, IAG is the root cause, and KGC is always influenced by others. Regarding the stocks from the remaining sectors, DIS, HST, and HD from consumer service, ORCL from technology, and CSX from transportation have large centrality.

By considering nonstationarity, 63 edges that were produced by the original PC algorithm were removed; 19 of them are within sectors, and 44 are between sectors. The causal modules of 37 out of 80 stocks change over time; most of them are from energy (10 out of 28), finance (7 out of 9), basic industry (6 out of 14), and consumer service (5 out of 7). It is reasonable that stocks, that are close to the root node, are more likely to be influenced by some external factors, such as changing economic environments and policies. These findings, such as stocks from finance are more likely to be causes of stocks in other sectors, are consistent with what we discovered from HK stock market reported above.

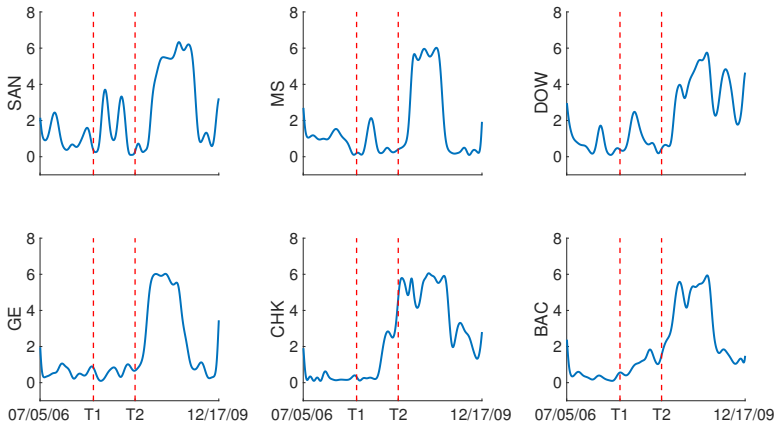


Figure 18: The estimated nonstationary driving force of six stock returns from 07/05/2006 $\sim$ 12/16/2009. The stocks SAN, MS, and DOW have obvious change points at 07/16/2007 (T_1) and 05/05/2008 (T_2) . The stocks GE, CHK, and BAC have change points only at 05/05/2008 (T_2). The change points match the critical time of financial crisis.

Figure 18 visualizes the estimated nonstationary driving forces of stocks SAN, MS, DOW, GE, CHK, and BAC. We can see that among these six stocks, SAN, MS, and DOW have obvious change points around 07/16/2007 (T_1) and 05/05/2008 (T_2), while the stocks GE, CHK, and BAC only have changes points around 05/05/2008 (T_2). We found that stocks, which have change points around both time points, usually have large centrality and influence a large number of other stocks. The change points match the critical time of the 2008 financial crisis - those in the TED spread and parts of the change points (T_2 and T_3) in HK stock data.

9. Discussion and Conclusions

This paper is concerned with causal discovery from heterogeneous/nonstationary data and visualization of changing causal modules. We assume a pseudo causal sufficiency condition, which states that all confounders can be represented by functions of domain or smooth functions of time index. We proposed (1) an enhanced constraint-based method for locating variables whose causal modules change and estimating the skeleton of the causal structure over observed variables, (2) a method for causal direction determination that takes advantage of the information carried by distribution shifts, and (3) a procedure for estimating a low-dimensional representation of changing causal modules. We then showed that distribution shifts also help for causal discovery when there are stationary confounders. In summary, to identify the causal structure, we take advantage of two types of independence constraints: (a) conditional independence (with constraint-based approach) and (b) independent changes of changing causal modules.

The performance evaluated on synthetic data sets showed largely improved F_1 score and precision with our proposed CD-NOD, compared to other methods for causal discovery. The results over task-fMRI data well revealed information flows between brain regions and how causal influences change across resting state and task states. The causal relationships learned from stock data (both HK and US) are consistent with background knowledge, e.g., stocks in finance and energy are causes to others, and the estimated nonstationary driving forces match the critical time of financial crisis around the year of 2008.

We showed that distribution shift contains useful information for causal discovery, since causal model provides a compact description of how the joint distribution changes. Moreover, it has been recently noticed that this view help understand and solve some machine learning problems, e.g., domain adaptation and prediction in nonstationary environments. Since distribution shifts in heterogeneous/nonstationary data are usually constrained - it might be due to the changes in the data-generating processes of only a few variables, and thus it is only necessary to adapt a small proportion of the joint distribution in domain adaptation problems (Zhang et al., 2013; Schölkopf et al., 2012).

There are several open questions that we aim to answer in future work. First, in the paper we assumed that causal directions do not flip despite distribution shifts. What if some causal directions also change across domains or over time? It is important to develop a general approach to detect such direction flips and do causal discovery, because of the ubiquity of feedback loops. Second, the issue of distribution shift may decrease the power of statistical (conditional) independence tests. It is essential to develop a reliable statistical test to mitigate this problem. Third, in practice, there may also exist other types of confounders, e.g., nonstationary confounders. We will extend our methods to cover general types of confounders.

Acknowledgments

We are grateful to the anonymous reviewers whose careful comments and suggestions helped improve this manuscript. We would like to acknowledge the support by National Institutes of Health under Contract No. NIH-1R01EB022858-01, FAIN-R01EB022858, NIH-1R01LM012087, NIH5U54HG008540-02, and FAIN-U54HG008540, and by the United States Air Force under Contract No. FA8650-17-C7715. The National Institutes of Health and the U.S. Air Force are not responsible for the views reported in this article. JZ's research was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

Appendix A. Proof of Theorem 1

Proof Before getting to the main argument, let us establish some implications of the SEMs in Eq. 2 and the assumptions in Section 3.1. Since the structure is assumed to be acyclic or recursive, according to Eq. 2, all variables V_i can be written as a function of $\{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}$ and $\{\epsilon_m\}_{m=1}^{n}$. As a consequence, the probability distribution of $\mathbf{V}$ at each value of C is determined by the distribution of $\epsilon_1, ..., \epsilon_n$, and the values of $\{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}$. In other words, $P(\mathbf{V}|C)$ is determined by $\prod_{i=1}^{n} P(\epsilon_i)$ (for

$\epsilon_1, ..., \epsilon_n$ are mutually independent), and $\{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$, where $P(\cdot)$ denotes the probability density or mass function. For any V_i, V_j, and $\mathbf{V}^{ij} \subseteq \{V_k \,|\, k \neq i, k \neq j\}$, because $P(V_i, V_j \,|\, \mathbf{V}^{ij}, C)$ is determined by $P(\mathbf{V}|C)$, it is also determined by $\prod_{i=1}^n P(\epsilon_i)$ and $\{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$. Since $\prod_{i=1}^n P(\epsilon_i)$ does not change with C, we have

$$\begin{aligned} &P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n \cup \{C\}) \\ =&P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n). \end{aligned} \tag{21}$$

That is,

$$C \perp\!\!\!\perp (V_i, V_j) \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n. \tag{22}$$

By the weak union property of conditional independence, it follows that

$$C \perp\!\!\!\perp V_j \,|\, \{V_i\} \cup \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n. \tag{23}$$

We are now ready to prove the theorem. Let V_i, V_j be any two variables in $\mathbf{V}$. First, suppose that V_i and V_j are not adjacent in G. Then they are not adjacent in G^{aug}, which recall is the graph that incorporates $\{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$. It follows that there is a set $\mathbf{V}^{ij} \subseteq \{V_k \,|\, k \neq i, k \neq j\}$ such that $\mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$ d-separates V_i from V_j. Since the joint distribution over $\mathbf{V} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$ is assumed to be Markov to G^{aug}, we have

$$V_i \perp\!\!\!\perp V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n. \tag{24}$$

Because all $g_l(c)$ and $\theta_m(C)$ are deterministic functions of C, we have $P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{C\}) = P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n \cup \{C\})$.

According to the properties of mutual information given in Madiman (2008) , Eqs. 24 and 22 imply $V_i \perp\!\!\!\perp (C, V_j) \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$. By the weak union property of conditional independence, it follows that $V_i \perp\!\!\!\perp V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n \cup \{C\}$. As all $g_l(C)$ and $\theta_m(C)$ are deterministic functions of C, it follows that $V_i \perp\!\!\!\perp V_j \,|\, \mathbf{V}^{ij} \cup \{C\}$. In other words, V_i and V_j are conditionally independent given a subset of $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$.

Conversely, suppose V_i and V_j are conditionally independent given a subset $\mathbf{S}$ of $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$. We show that V_i and V_j are not adjacent in G, or equivalently, that they are not adjacent in G^{aug}. There are two possible cases to consider:

- Suppose $\mathbf{S}$ does not contain C. Then since the joint distribution over $\mathbf{V} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$ is assumed to be faithful to G^{aug}, V_i and V_j are not adjacent in G^{aug}, and hence not adjacent in G.
- Otherwise, $\mathbf{S} = \mathbf{V}^{ij} \cup \{C\}$ for some $\mathbf{V}^{ij} \subseteq \{V_k \,|\, k \neq i, k \neq j\}$. That is,

$$\begin{aligned} &V_i \perp\!\!\!\perp V_j \,|\, \mathbf{V}^{ij} \cup \{C\}, \text{ or} \\ &P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{C\}) = P(V_i \,|\, \mathbf{V}^{ij} \cup \{C\}) P(V_j \,|\, \mathbf{V}^{ij} \cup \{C\}). \end{aligned} \tag{25}$$

  According to Eq. 21, and also noting that $\{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n$ is a deterministic function of C, we have

$$P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{C\}) = P(V_i, V_j \,|\, \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^L \cup \{\theta_m(C)\}_{m=1}^n), \tag{26}$$

which also implies

$$P(V_i \mid \mathbf{V}^{ij} \cup \{C\}) = P(V_i \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}), \tag{27}$$

$$P(V_j \mid \mathbf{V}^{ij} \cup \{C\}) = P(V_j \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}). \tag{28}$$

Substituting Eqs. 26 - 28 into Eq. 25 gives

$$\begin{aligned} &P(V_i, V_j \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}) \\ =&P(V_i \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}) P(V_j \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}). \end{aligned} \tag{29}$$

That is,

$$V_i \perp\!\!\!\perp V_j \mid \mathbf{V}^{ij} \cup \{g_l(C)\}_{l=1}^{L} \cup \{\theta_m(C)\}_{m=1}^{n}.$$

Again, by the Faithfulness assumption on G^{aug}, this implies that V_i and V_j are not adjacent in G^{aug} and hence are not adjacent in G.

Therefore, V_i are V_j are not adjacent in G if and only if they are conditionally independent given some subset of $\{V_k \mid k \neq i, k \neq j\} \cup \{C\}$.

■

Appendix B. Proof of Proposition 1

Proof

$$\begin{aligned} &\tilde{\mu}_{\underline{Y},X|C=c_n} \triangleq \mathbb{E}_{(Y,X)\sim\tilde{P}(\underline{Y},X \mid C=c_n)}[\phi(Y) \otimes \phi(X)] \\ =&\tilde{\mathcal{C}}_{(Y,X),C}\mathcal{C}_{C,C}^{-1}\phi(c_n) \\ =&\mathbb{E}_{(Y,X,C)\sim\tilde{P}(\underline{Y},X,C)}[\phi(Y) \otimes \phi(X) \otimes \phi(C)]\mathcal{C}_{C,C}^{-1}\phi(c_n) \\ =&\mathbb{E}_{(X,C)\sim\tilde{P}(X,C)}\{\mathbb{E}_{Y\sim P(Y|X,C)}[\phi(Y)] \otimes \phi(X) \otimes \phi(C)\}\mathcal{C}_{C,C}^{-1}\phi(c_n) \\ =&\mathcal{C}_{Y,(X,C)}\mathcal{C}_{(X,C),(X,C)}^{-1} \cdot \\ &\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{[\phi(X) \otimes \phi(C)] \otimes \phi(X) \otimes \phi(C)\}\mathcal{C}_{C,C}^{-1}\phi(c_n) \end{aligned} \tag{30}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{[\phi(X) \otimes \phi(C)] \otimes \phi(X) \otimes \phi(C)\}\mathcal{C}_{C,C}^{-1}\phi(c_n) \\ =&\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{[\phi(X) \otimes \phi(C)] \otimes [\phi(X) \cdot \phi^\intercal(C)\mathcal{C}_{C,C}^{-1}\phi(c_n)]\} \\ =&\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{\phi(X) \otimes [\phi(C) \cdot (\phi(X)\phi^\intercal(C)\mathcal{C}_{C,C}^{-1}\phi(c_n))^\intercal]\} \qquad (31) \\ =&\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{\phi(X) \otimes [\phi(C) \cdot \phi^\intercal(c_n)\mathcal{C}_{C,C}^{-1}\phi(C)\phi^\intercal(X)]\} \\ =&\mathbb{E}_{X\sim P(X)}\mathbb{E}_{C\sim P(C)}\{\phi(X) \otimes [\mathcal{C}_{C,C}\mathcal{C}_{C,C}^{-1}\phi(c_n)\phi^\intercal(X)]\} \qquad (32) \\ =&\mathbb{E}_{X\sim P(X)}\{\phi(X) \otimes \phi(c_n) \otimes \phi(X)\} \end{aligned}$$

where (31) holds because tensor product is associative, and (32) holds because $\phi^\intercal(c_n)\mathcal{C}_{C,C}^{-1}\phi(C)$ is a scaler, implying $\phi^\intercal(c_n)\mathcal{C}_{C,C}^{-1}\phi(C) = [\phi^\intercal(c_n)\mathcal{C}_{C,C}^{-1}\phi(C)]^\intercal = \phi^\intercal(C)\mathcal{C}_{C,C}^{-1}\phi(c_n)$.

Therefore, equation (30) becomes

$$\tilde{\mu}_{\underline{Y},X|C=c_n} = \mathcal{C}_{Y,(X,C)}\mathcal{C}^{-1}_{(X,C),(X,C)}\mathbb{E}_{X\sim P(X)}\{\phi(X)\otimes\phi(c_n)\otimes\phi(X)\}.$$

Let $\phi^{\otimes}(X,C) := \phi(X)\otimes\phi(C)$, $\mathbf{\Phi_y} := [\phi(y_1),...,\phi(y_N)]$, $\mathbf{\Phi_x} := [\phi(x_1),...,\phi(x_N)]$, $\mathbf{\Phi_{x,c}} := [\phi^{\otimes}(x_1,c_1),...,\phi^{\otimes}(x_N,c_n)]$, $\mathbf{\Phi}_{\mathbf{x},c_n} := [\phi^{\otimes}(x_1,c_n),...,\phi^{\otimes}(x_N,c_n)]$. $\tilde{\mu}_{\underline{Y},X|C=c_n}$ can be estimated as

$$\begin{aligned}\hat{\tilde{\mu}}_{\underline{Y},X|C=c_n} &= \frac{1}{N}\mathbf{\Phi_y}\mathbf{\Phi}^{\intercal}_{\mathbf{x},\mathbf{c}}(\frac{1}{N}\mathbf{\Phi_{x,c}}\mathbf{\Phi}^{\intercal}_{\mathbf{x},\mathbf{c}}+\lambda I)^{-1}(\frac{1}{n}\mathbf{\Phi}_{\mathbf{x},c_n}\mathbf{\Phi}^{\intercal}_{\mathbf{x}})\\ &= \frac{1}{N}\mathbf{\Phi_y}(\mathbf{K_x}\odot\mathbf{K_c}+\lambda I)^{-1}\mathrm{diag}(\mathbf{k}_{\mathbf{c},c_n})\mathbf{K_x}\mathbf{\Phi}^{\intercal}_{\mathbf{x}}.\end{aligned}$$

■

Appendix C. Proof of Theorem 2

Proof Let V_i and V_j be an unoriented pair of adjacent variables.

If condition (1) is satisfied, then there are two possibilities:

- V_j is the collider, and there exists another variable V_k which influences V_j and does not have an edge with V_i, so that $V_i \to V_j \leftarrow V_k$. In such a case, $V_i \perp\!\!\!\perp V_k|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $\mathbf{S}\wedge V_j = \emptyset$.
- V_i is the collider, and there exists another variable V_k which influences V_i and does not have an edge with V_j, so that $V_j \to V_i \leftarrow V_k$. In such a case, $V_j \perp\!\!\!\perp V_k|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $\mathbf{S}\wedge V_i = \emptyset$.

Thus, we can identify the direction between V_i and V_j depending on whether $V_i \perp\!\!\!\perp V_k|\mathbf{S}$ or $V_j \perp\!\!\!\perp V_{k'}|\mathbf{S}'$.

If condition (2) is satisfied, and there is only one change influences V_i or V_j. Then we have the following two cases:

- The change influences V_j, i.e., $V_i - V_j \leftarrow C$:
  - if $V_i \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $\mathbf{S}\wedge V_j = \emptyset$, orient $V_i \to V_j$;
  - if $V_i \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $V_j\subseteq\mathbf{S}$, orient $V_i \leftarrow V_j$.
- The change influences V_i, i.e., $V_j - V_i \leftarrow C$:
  - if $V_j \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $\mathbf{S}\wedge V_i = \emptyset$, orient $V_j \to V_i$;
  - if $V_j \perp\!\!\!\perp C|\mathbf{S}$, with $\mathbf{S}\subseteq\mathbf{V}$ and $V_i\subseteq\mathbf{S}$, orient $V_j \leftarrow V_i$.

If condition (2) is satisfied, and there are independent changes to both V_i and V_j. Then we have the following two cases:

- if $P(V_i) \perp\!\!\!\perp P(V_j|V_i,\mathbf{Z})$ and $P(V_j) \not\perp\!\!\!\perp P(V_i|V_j,\mathbf{Z})$, where $\mathbf{Z}$ is the deconfounding set of (V_i,V_j), then orient $V_i \to V_j$;
- if $P(V_j) \perp\!\!\!\perp P(V_i|V_j,\mathbf{Z})$ and $P(V_i) \not\perp\!\!\!\perp P(V_j|V_i,\mathbf{Z})$, where $\mathbf{Z}$ is the deconfounding set of (V_i,V_j), then orient $V_j \to V_i$.

If condition (3) is satisfied, then we have the following two cases:

- There is an edge incident to V_j but not to V_i, i.e., $V_i - V_j \leftarrow V_l$:

  - if $V_i \perp\!\!\!\perp V_l | \mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $\mathbf{S} \wedge V_j = \emptyset$, orient $V_i \rightarrow V_j$;
  - if $V_i \perp\!\!\!\perp V_l | \mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $V_j \subseteq \mathbf{S}$, orient $V_i \leftarrow V_j$.
- There is an edge incident to V_i but not to V_j, i.e., $V_j - V_i \leftarrow V_l$:
  - if $V_j \perp\!\!\!\perp V_l | \mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $\mathbf{S} \wedge V_i = \emptyset$, orient $V_j \rightarrow V_i$;
  - if $V_j \perp\!\!\!\perp V_l | \mathbf{S}$, with $\mathbf{S} \subseteq \mathbf{V}$ and $V_i \subseteq \mathbf{S}$, orient $V_j \leftarrow V_i$.

Therefore, it is easy to see that the direction between V_i and V_j is identifiable, if at least one of the conditions is satisfied. ■

References

R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

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

P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

A. E. Ghassami, N. Kiyavash, B. Huang, and K. Zhang. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6266–6276, 2018.

M. Gong*, K. Zhang*, D. Tao, P. Geiger, and B. Schölkopf. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning (ICML)*, pages 1898–1906, 2015.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Neural Information Processing Systems (NIPS)*, pages 585–592, 2008.

M. Havlicek, K.J. Friston, J. Jan, M. Brazdil, and V.D. Calhoun. Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *Neuroimage*, 56:2109–2128, 2011.

Y. He and Z. Geng. Causal network learning from multiple interventions of unknown manipulated targets. In *arXiv preprint arXiv: 1610.08611*, 2016.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. *Innovations in Machine Learning*, 194:1–28, 2006.

K. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

P. O. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

B. Huang, K. Zhang, and B. Schölkopf. Identification of time-dependent causal model: A Gaussian process treatment. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3561–3568, 2015.

B. Huang, K. Zhang, J. Zhang, Sanchez-Romero, C. R., Glymour, and B. Schölkopf. Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE International Conference on Data Mining (ICDM)*, pages 913–918, 2017.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1551–1560, 2018.

B. Huang, K. Zhang, M. Gong, and C. Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International Conference on Machine Learning (ICML)*, 2019a.

B. Huang, K. Zhang, P. Xie, M. Gong, E. Xing, and C. Glymour. Specific and shared causal relation modeling and mechanism-based clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

B. Huang, K. Zhang, M. Gong, and C. Glymour. Causal discovery from non-identical variable sets. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, pages 1709–1731, 2010.

E. Kummerfeld and D. Danks. Tracking time-varying graphical structure. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1205–1213, 2013.

M. Madiman. On the entropy of sums. In *IEEE Information Theory Workshop (ITW)*, pages 303–307, 2008.

C. Meek. Strong completeness and faithfulness in bayesian networks. In *Uncertainty in Artificial Intelligence (UAI)*, pages 411–419, 1995.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

B. Schölkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, 2012.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

L. Song, M. Kolar, and E. Xing. Time-varying dynamic Bayesian networks. In *Advances in Neural Information Processing systems (NIPS)*, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Spring-Verlag Lectures in Statistics, 1993.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.

M. Talih and N. Hengartner. Structural learning with time-varying components: Tracking the cross-section of financial time series. *Journal of the Royal Statistical Society - Series B*, 67(3):321–341, 2005.

J. Tian and J. Pearl. Causal discovery from changes: a Bayesian approach. In *Uncertainty in Artificial Intelligence (UAI)*, pages 512–521, 2001.

X. Wang, R. Hutchinson, and T. M. Mitchell. Training fMRI classifiers to detect cognitive states across multiple human subjects. In *Neural Information Processing Systems (NIPS)*, pages 709–716, 2004.

E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 4(2):535–566, 2010.

J. Zhang and P. Spirtes. The three faces of faithfulness. *Synthese*, 2016.

K. Zhang and L. Chan. Extensions of ICA for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing (ICONIP)*, 2006.

K. Zhang and A. Hyvärinen. Acyclic causality discovery with additive noise: An information-theoretical perspective. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2009a.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Uncertainty in Artificial Intelligence*, 2009b.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Uncertainty in Artificial Intelligence (UAI)*, 2011.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, 2013.

K. Zhang, B. Huang, J. Zhang, B. Schölkopf, and C. Glymour. Discovery and visualization of nonstationary causal models. In *arXiv preprint arXiv: 1509.08056*, 2015a.

K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 7(2), 2015b.

K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.