# Causal Discovery with Heterogeneous Observational Data


## Abstract

We consider the problem of causal discovery (structure learning) from heterogeneous observational data. Most existing methods assume a homogeneous sampling scheme, which may lead to misleading conclusions when violated. We propose a novel approach that exploits data heterogeneity to infer possibly cyclic causal structures from causally insufficient systems. The core idea is to model the direct causal effects as functions of exogenous covariates that help explain data/sampling heterogeneity. We investigate the structure identifiability properties of the proposed model. Structure learning is carried out in a fully Bayesian fashion, which provides natural uncertainty quantification. We demonstrate its utility through extensive simulations and two real-world applications.


## 1 INTRODUCTION

Causal discovery is a central task in various fields including social science, artificial intelligence, and systems biology. While randomized controlled trials are the gold standard to establish causality, they can be too costly, unethical, or impossible to carry out. For example, recovering gene regulatory networks through gene knockout would be too expensive to scale whereas observational genomic data are considerably easier to obtain with next-generation sequencing technologies and have become widely available. Many causal discovery methods, therefore, attempt to discover causality from purely observational data.

**Related work** One prominent approach in presenting and learning causality is to use a structural equation model (SEM) and the associated causal graph [Pearl, 1998]. The recursive linear Gaussian SEM is among the most popular ones although the associated causal directed acyclic graph is only identifiable up to Markov equivalence classes [Verma and Pearl, 1990]. In order to uniquely identify causal structures with observational data, additional distributional assumptions have been made in prior works including the linear non-Gaussian model [Shimizu et al., 2006], the nonlinear additive noise model [Hoyer et al., 2008a], and the linear Gaussian model with equal error variances [Peters and Bühlmann, 2014]. A common thread of these methods is that they assume that the causal graph is acyclic and there are no unmeasured confounders (also known as causal sufficiency). However, directed cycles and confounders are very common in practice. For example, feedback loops (directed cycles) are common regulatory motifs in biological signaling systems [Brandman and Meyer, 2008]. As for confounders, gene regulation is known to be affected by many factors such as epigenetic modification [Portela and Esteller, 2010], which may not be measured together with gene expressions.

To allow for cycles, non-recursive SEMs have been developed and proven to be identifiable for linear non-Gaussian models [Lacerda et al., 2008] and nonlinear additive noise models [Mooij et al., 2011]. In the presence of unmeasured confounders, linear non-Gaussian SEMs have received lots of attention: various models have been proposed and shown to be structurally identifiable under the respective confounding assumptions [Hoyer et al., 2008b, Chen and Chan, 2012, Shimizu and Bollen, 2014, Salehkaleybar et al., 2020]. Nevertheless, none of the aforementioned methods explicitly deal with and provide identifiability guarantees for graphs with both cycles and confounders. Although Hyttinen et al. [2012] and Forré and Mooij [2018] provided learning algorithms for general SEMs (allowing cycles, confounders, and nonlinearity), full structure can only be recovered with interventional data, which is quite different from the purely observational setting considered in this paper.

Furthermore, all the aforementioned methods assume independent and identically distributed (iid) observations, which may be violated in many applications. For example, cancer is known to be a genetically heterogeneous disease and,

*Submitted to the 38th Conference on Uncertainty in Artificial Intelligence* (UAI 2022). **To be used for reviewing only**.

therefore, cancer genomic data exhibit great heterogeneity. Methods that ignore such heterogeneity can perform poorly as we will see in our experiments. Recently, Ni et al. [2019], Huang et al. [2020], Saeed et al. [2020] explicitly addressed the heterogeneity issue by incorporating covariates or using a latent mixture model but their models are acyclic and do not account for unmeasured confounders. The method proposed by Peters et al. [2016] is able to identify the causal relationships in heterogeneous data if the causal effects are invariant across environments. By contrast, we assume the causal mechanism varies with the environment, which is motivated by biological systems. For instance, critical gene pathways have been known to change during the course of cancer development [Moustakas and Heldin, 2007, Huang et al., 2009] and the causal effects between two genes/proteins could be vastly different across patients due to different cancer subtypes/stages; hence, in this case, causal variance rather than invariance may be a more appropriate assumption. Mooij et al. [2020] proposed a flexible joint causal inference (JCI) framework, which allows the causal mechanism to vary. However, it did not perform well in our specific real data example (see Section 4.2).

In this paper, we propose a novel method for **C**ausal discovery with **H**eterogeneous **O**bservational **D**ata (CHOD). Importantly, we do not restrict our model to be acyclic and do not assume causal sufficiency. By exploiting the data heterogeneity via exogenous covariates, we provide sufficient conditions under which CHOD is structurally identifiable in (i) causally insufficient bivariate cyclic graphs, (ii) causally insufficient multivariate acyclic graphs, and (iii) causally sufficient multivariate cyclic graphs. Our method is among the first model-based causal discovery methods to identify causal graphs with both cycles and confounders in purely observational settings without prior domain knowledge. Extensive simulation experiments and two real-world applications support the utility of our method and demonstrate its superiority in handling heterogeneous data through comparison with state-of-the-art alternatives.

## 2 PRELIMINARIES OF CAUSAL DISCOVERY

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a $p$-dimensional random vector. We represent the causal structure as well as the joint observational distribution of $\boldsymbol{X}$ by a linear SEM, $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{E}$, with direct causal effects $\boldsymbol{B} = [b_{j\ell}] \in \mathbb{R}^{p\times p}$ and random noises $\boldsymbol{E} = [e_j] \in \mathbb{R}^p$. If $b_{j\ell} \neq 0$, then $X_\ell$ is a direct cause of $X_j$. We assume $\boldsymbol{E}$ to be centered Gaussian with covariance $\boldsymbol{S} = [\sigma_{j\ell}]$. When there are no unmeasured confounders (i.e., hidden common causes), the noises are independent of each other and hence $\boldsymbol{S}$ is diagonal. However, the presence of confounders would correlate the noises, making the off-diagonal elements of $\boldsymbol{S}$ non-zero and resulting in a causally insufficient system. To see that, suppose we explicitly model unmeasured Gaussian confounders $\boldsymbol{L}$ via $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\Gamma}\boldsymbol{L} + \boldsymbol{E}'$. Then marginalizing out $\boldsymbol{L}$ leads to $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{E}$ where $\boldsymbol{S} = \mathrm{cov}(\boldsymbol{E}) = \boldsymbol{\Gamma}\mathrm{cov}(\boldsymbol{L})\boldsymbol{\Gamma}^T + \mathrm{cov}(\boldsymbol{E}')$. Therefore, the non-zero off-diagonal entries of $\boldsymbol{S}$ implicitly account for unmeasured confounders.

We use a *mixed graph* $\mathcal{G}_M = (V, E_B, E_D)$ to represent the causal relationships embedded in the SEM, where $V = \{1, \ldots, p\}$ is the set of *nodes* representing $\boldsymbol{X}$, $E_B$ is the set of *bidirected edges*, and $E_D$ is the set of *directed edges*; see Figure 1 for a few examples. There is a bidirected edge $\ell \leftrightarrow j$ if $\sigma_{j\ell} \neq 0$, and a directed edge $\ell \to j$ if $b_{j\ell} \neq 0$. In the former case, $X_j$ and $X_\ell$ are confounded by at least one hidden common cause, while in the latter case, $X_\ell$ is a direct cause of $X_j$. The graph is acyclic if there does not exist a directed path $k_1 \to k_2 \to \ldots \to k_\ell \to k_1$ that returns a node to itself, otherwise it is called cyclic. Our goal is to identify the edge-induced subgraph $\mathcal{G} = (V, E_D)$ with direct causal relationships $E_D$ among the observed variables $\boldsymbol{X}$.

## 3 METHOD

### 3.1 PROPOSED MODEL

Our key idea to discover causality is to take advantage of the data heterogeneity, which we assume can be explained by some exogenous covariates $\boldsymbol{Z} \in \mathbb{R}^q$. The exogenous covariates may be observed (e.g., biomarkers in cancer genomic data) or latent. In the latter case, one can impute the latent covariates by various embedding methods such as t-SNE [van der Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018]. Alternatively, latent covariates can be learned simultaneously with our model. For ease of exposition, we first focus our discussion on the case where the exogenous covariate is given (either observed or imputed) and univariate (i.e., $q = 1$), and later briefly discuss the extension to multivariate latent covariates. Specifically, given $Z$, we model $\boldsymbol{X}$ as a conditionally linear Gaussian SEM,

$$\boldsymbol{X} = \boldsymbol{B}(Z)\boldsymbol{X} + \boldsymbol{E}, \;\; \boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{S}), \tag{1}$$

where $\boldsymbol{B}(Z) = [b_{j\ell}(Z)] : \mathbb{R} \mapsto \mathbb{R}^{p\times p}$ is a matrix-valued function of $Z$, which characterizes the changes of the direct causal effects with respect to $Z$. Because each observation potentially has a different value of covariate $Z$, the direct causal effects $\boldsymbol{B}(Z)$ are heterogeneous and observation-specific. Model (1) implies the conditional distribution of $\boldsymbol{X}$ given $Z$,

$$\mathbb{P}(\boldsymbol{X}|Z, \boldsymbol{B}, \boldsymbol{S}) = \det(\boldsymbol{I} - \boldsymbol{B}(Z))N((\boldsymbol{I} - \boldsymbol{B}(Z))\boldsymbol{X}|\boldsymbol{0}, \boldsymbol{S}).$$

When $\boldsymbol{B}(Z)$ is constant in $Z$, model (1) is reduced to an ordinary linear Gaussian SEM and hence its underlying

causal graph $\mathcal{G}$ is not identifiable. However, as we will show later, the causal graph of model (1) is in general identifiable.

**Relation to existing methods.** While most existing causal discovery methods that apply to heterogeneous data assume the exogenous covariate to be discrete and finite (i.e., multiple contexts, domains, or experimental conditions), under our framework the exogenous covariate $Z$ can be either continuous or discrete. To match the case of our real-data application (see Section 4.2) and emphasize the advantage of CHOD, we present $Z$ as a continuous covariate in this paper. With continuous $Z$, the proposed method is particularly useful when the data are heterogeneous but there are no clear predefined discrete groups. If the covariate is categorical, the proposed model can be thought of as multi-domain/group-specific graphical models (see, for example, Yajima et al. [2015], Ghassami et al. [2018], Ni et al. [2018]) by viewing the categorical covariate as the domain or group indicator. The proposed model is also reminiscent of causal models with soft interventions by viewing $Z$ as intervention that modifies causal effects; however, the key difference is that our model does not assume the knowledge of the interventional targets (i.e., $Z$ could affect all causal effects) and the interventions are conducted by nature rather than humans. In summary, our model explicitly accounts for the heterogeneity of data generating mechanism via the observation-specific direct causal effects $\boldsymbol{B}(Z)$, which vary smoothly with covariate $Z$.

In the regression context, the varying-coefficient model serves as an important generalization of linear model. As introduced by Hastie and Tibshirani [1993], the class of varying-coefficient models ties together many important structured regression models such as additive models and dynamic linear models into one common framework. Likewise, our proposed model is a natural extension of linear SEMs, which allows varying causal effects. One important ingredient exploited in this paper is that the adoption of varying causal effects helps identify the causal structure. Note that for simplicity, we have assumed linearity and Gaussianity in the current formulation. In addition to efficient computation and causal effect estimation (discussed briefly in Section S4 of the Supplementary Materials), this simple setup allows us to convey the main idea that heterogeneity alone is enough to enable causal identification.

### 3.2 CAUSAL STRUCTURE IDENTIFIABILITY

For model-based causal discovery methods, the non-identifiability issue can be seen from the distribution equivalence point of view. Two CHOD models parameterized by $(\boldsymbol{B}, \boldsymbol{S})$ and $(\boldsymbol{B}', \boldsymbol{S}')$ are said to be *distribution equivalent* if for any values of $(\boldsymbol{B}, \boldsymbol{S})$ there exist values of $(\boldsymbol{B}', \boldsymbol{S}')$ such that $\mathbb{P}(\boldsymbol{X}|Z, \boldsymbol{B}, \boldsymbol{S}) = \mathbb{P}(\boldsymbol{X}|Z, \boldsymbol{B}', \boldsymbol{S}')$ for all $\boldsymbol{X}$. Clearly, distribution equivalent models cannot be distinguished from each other by examining their observational distributions. The causal structure is said to be *identifiable* if there do not exist two distribution equivalent causal models such that $\mathcal{G} \neq \mathcal{G}'$.

Throughout, we make causal Markov assumption [Richardson, 1996], i.e., the probability distribution $\mathbb{P}$ respects the Markov property of the causal graph $\mathcal{G}$. Before stating our main results, we first provide an intuition on how the proposed CHOD is identifiable using a toy example. Consider the bivariate graphs shown in Figure 1. We can distinguish graphs (a)–(b) from graphs (c)–(f) because the marginal variance of $X_2$ is independent of $Z$ in graphs (a)–(b) but depends on $Z$ in graphs (c)–(f) through the causal effect $b_{21}(Z)$: $X_1 \to X_2$. Likewise, we can separate graphs (c)–(d) from graphs (e)–(f) by examining the marginal variance of $X_1$ which depends on $Z$ through $X_2 \to X_1$. We may not distinguish (c) from (d) or (e) from (f), but the direct causal relationship between $X_1$ and $X_2$ is determined in either case.

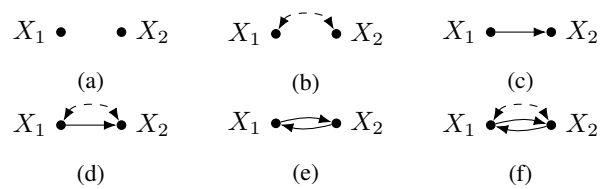
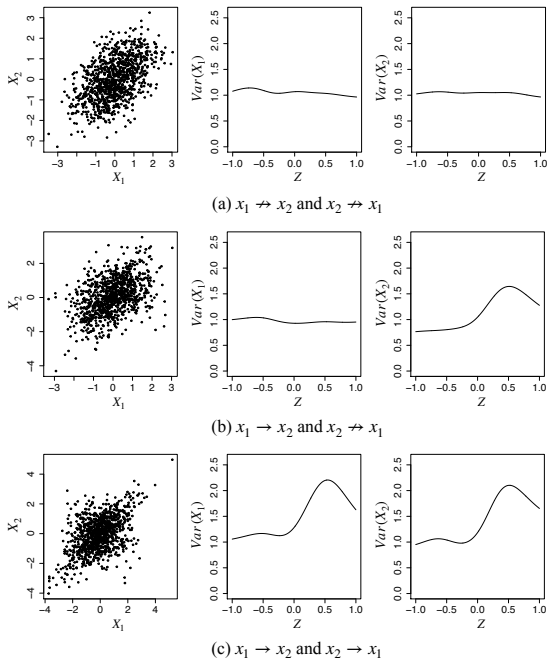



Figure 1: Mixed graphs. Solid arrows are causal effects and dashed bidirected arrows are confounding effects.

We further illustrate the identifiability with simulated data from graphs (b), (d), and (f) in Figure 1. Specifically, the exogenous covariates were generated uniformly. Under each graph, the non-zero elements of $\boldsymbol{B}(Z)$ were assumed to be $0.5\sin(\pi Z)$. We set the noise variances to 1 and the correlation coefficients to 0.5 to have confounding effects. The $n = 1000$ data points as well as the marginal variances estimated by kernel methods of the two nodes as functions of $Z$ are depicted in Figure 2 for these 3 cases, from which the causal relationships between $X_1$ and $X_2$ are intuitively identifiable in the presence of both confounders and cycles: in Figure 2(a), both Var$(X_1)$ and Var$(X_2)$ are constant in $Z$ indicating no direct causal link; in Figure 2(b), Var$(X_1)$ is constant but Var$(X_2)$ is not constant in $Z$ indicating a direct causal link $X_1 \to X_2$; and in Figure 2(c), neither Var$(X_1)$ nor Var$(X_2)$ is constant in $Z$ indicating a cyclic causal link $X_1 \rightleftarrows X_2$.

Now we present the identifiability theories. We first show that in the bivariate case the CHOD models are not distribution equivalent and therefore their causal graphs are identifiable.

**Theorem 1** (Causally Insufficient Bivariate Cyclic Graphs)**.** *Consider bivariate CHOD models with direct causal effects* $[b_{12}(Z), b_{21}(Z)]$ *and* $[b'_{12}(Z), b'_{21}(Z)]$*, respectively. Assume* $b_{j\ell}(Z)$ *and* $b'_{j\ell}(Z)$ *are either zero or non-constant*




(a) $x_1 \nrightarrow x_2$ and $x_2 \nrightarrow x_1$

(b) $x_1 \rightarrow x_2$ and $x_2 \nrightarrow x_1$

(c) $x_1 \rightarrow x_2$ and $x_2 \rightarrow x_1$

Figure 2: Illustration with a bivariate toy example.

*functions for all $j \neq \ell \in \{1, 2\}$. Then if the two CHOD models are distribution equivalent, we must have $\mathcal{G} = \mathcal{G}'$.*

All proofs are provided in Section S1 of the Supplementary Materials. The assumption that $b_{j\ell}(Z)$ is either zero or a non-constant function is not surprising because if non-zero $b_{j\ell}(Z)$ is constant in $Z$, then the proposed model is reduced to an ordinary linear Gaussian SEM, which is known to be non-identifiable. Loosely speaking, Theorem 1 states that CHOD is identifiable if $Z$ can help explain the heterogeneity of the data generating mechanism.

Next, we provide sufficient conditions for causally insufficient multivariate acyclic systems and causally sufficient multivariate cyclic systems to be identifiable, and leave the theoretical investigation of causally insufficient multivariate cyclic systems as future work. Denote $pa(j) = \{\ell : \ell \rightarrow j \in E_D\}$ as the set of direct causes and $ds(j) = \{\ell : \ell \leftrightarrow \cdots \leftrightarrow j\}$ as the set of nodes connected to $j$ through bidirected arrows.

**Theorem 2** (Causally Insufficient Multivariate Acyclic Graphs)**.** *Consider the CHOD model in* (1) *restricted to acyclic causal graphs. Assume without loss of generality $(1, \ldots, p)$ is a true causal ordering (i.e., $\ell \not\rightarrow j$ if $\ell > j$). If for any node $j$, and any set $S = \{1, \ldots, m\}$ such that $pa(j) \not\subset S$, we have $\mathrm{Var}(X_j|\boldsymbol{X}_S)$ is a non-constant function of the covariate $Z$, then the causal ordering is identifiable. Moreover, if $pa(j) \cap ds(j) = \emptyset, \forall j$, then the causal graph is identifiable.*

The proof of Theorem 2 first identifies an ordering by recursively finding root variables in acyclic graphs and then identifies the graph structure given the ordering. The assumption on $\mathrm{Var}(X_j|\boldsymbol{X}_S)$ means that the heterogeneous causal effects do not accidentally become constant in any paths, which is similar in spirit to the causal faithfulness assumption. See Section S1.2 of the Supplementary Materials for more discussions.

We have presented our theorems in their strongest forms, i.e., full structure identifiability. If some of the causal effects do not vary with $Z$, then their identification is not always guaranteed (they may still be identifiable in some graphs via v-structure and Meek rules). This is similar to other causal models. For example, in additive noise models, all causal effects have to be nonlinear for full identification. Those linear causal effects have to rely on v-structure and Meek rules to achieve identification as in our method. In the linear non-Gaussian acyclic model, all but one noises have to be non-Gaussian. Like our model, violation of these assumptions would lead to partial structure identification. We would like to point out though, under our proposed Bayesian learning framework discussed in Section 3.3, we can assess the credibility of inferred edges via posterior inference: edges that have nearly constant causal effects (e.g., if 95% credible bands of $b_{j\ell}(Z)$, which can be computed from Monte Carlo samples, cover constant functions) are deemed less reliable.

Unlike bivariate graphs, the identifiability results of multivariate cyclic graphs for purely observational data are sparse in the literature with few exception [Lacerda et al., 2008], which assumes causal sufficiency and disjoint cycles. In the following theorem, we also make the same assumptions.

**Theorem 3** (Causally Sufficient Multivariate Cyclic Graphs)**.** *Consider the CHOD model* (1)*. Assume there are no unmeasured confounders and all cycles are disjoint. The causal graph is generally identifiable*[1].

Theorems 1–3 assume $Z$ to be univariate and known (observed or imputed). When $\boldsymbol{Z}$ is multivariate and unknown, it can be inferred jointly with the causal graph. We provide its identifiability result below and briefly discuss its implementation in the Section S2.1 of the Supplementary Materials.

**Proposition 1** (Multivariate Latent Exogenous Covariates)**.** *Assume the vector $\boldsymbol{m}(\boldsymbol{Z})$ that stacks the non-zero elements of $\boldsymbol{B}(\boldsymbol{Z})$ is continuous and injective, and $(\boldsymbol{m}, \boldsymbol{S}) \mapsto \mathbb{P}(\boldsymbol{X}|\boldsymbol{m}, \boldsymbol{S})$ is continuous and injective in $\boldsymbol{m}$ given $\mathcal{G}$. Then the latent exogenous covariates are identifiable up to a monotone transformation.*

Proposition 1 shows that the *relative* position of the latent covariates can be identified, which is useful in sorting observations (see many prominent examples of trajectory inference

[1]That is, it is identifiable unless a peculiar condition holds. Because of space, we discuss that condition ($\star$) in the Section S1.3 of the Supplementary Materials.

in single-cell genomic studies [Saelens et al., 2019]). It can be also viewed as an embedding and dimension reduction tool wrapped in a causal model because the dimension of $\boldsymbol{Z}$ is typically much smaller than $\boldsymbol{X}$. The condition of Proposition 1 that assumes $\boldsymbol{m}(\boldsymbol{Z})$ to be injective as a vector-valued function should not be interpreted as a requirement that each individual function has to be injective. For example, if $b_{12}(Z) = (Z+1)^2$ and $b_{21}(Z) = Z^2$, neither is injective but the resulting $\boldsymbol{m}(Z) = [b_{12}(Z), b_{21}(Z)]$ is injective. Given the causal structure, the second requirement on the distribution $\mathbb{P}$ is equivalent to the identifiability of causal effects or model parameters. The causal effect identification itself is an interesting but challenging task. For linear Gaussian SEMs, it is well-known that causal effects are identifiable without confounders. With confounders, Drton et al. [2011] showed that the acyclic mixed graph needs to be a simple graph. As is evident from the proofs of Theorems 2 and 3, the causal effects under the corresponding assumptions are indeed identifiable for our model. In a related work, Salehkaleybar et al. [2020] showed that the causal effects in the presence of latent confounders are identifiable with mild structure assumptions in the non-Gaussian setting. This paper focuses on investigating causal structure identifiability; establishing causal effect identifiability theory for the causally insufficient multivariate cyclic graphs will be an interesting future work.

### 3.3 BAYESIAN STRUCTURE LEARNING

We learn the causal structure through a Bayesian approach by assigning priors on the space of graphs and model parameters. We model the direct causal effects by cubic B-splines with evenly spaced knots $b_{j\ell}(Z) = \sum_{k=1}^{K} \beta_{j\ell k}\phi_k(Z)$, where $\{\phi_k(Z)\}_{k=1}^{K}$ is the set of spline basis. To encourage graph sparsity, a spike-and-slab prior is assigned to the vector $\boldsymbol{\beta}_{j\ell} = (\beta_{j\ell 1}, \ldots, \beta_{j\ell K})^T$,

$$\mathbb{P}(\boldsymbol{\beta}_{j\ell}|r_{j\ell}, \tau) = (1 - r_{j\ell})\delta_{\mathbf{0}}(\boldsymbol{\beta}_{j\ell}) + r_{j\ell}N(\boldsymbol{\beta}_{j\ell}|\mathbf{0}, \tau\boldsymbol{I}),$$

where $\delta_{\mathbf{0}}(\cdot)$ is a point mass at vector zero and $r_{j\ell}$ is a binary edge indicator. By construction, $r_{j\ell} = 0$ if and only if $\boldsymbol{\beta}_{j\ell} = \mathbf{0}$ (equivalently, $\ell \not\to j$ and $b_{j\ell}(Z) \equiv 0$). We assume independent beta-Bernoulli priors with $r_{j\ell} \sim \mathbb{P}(r_{j\ell}|\pi) = \text{Bernoulli}(r_{j\ell}|\pi)$ and $\pi \sim \mathbb{P}(\pi) = \text{beta}(\pi|a, b)$. We place conjugate inverse-gamma prior on $\tau \sim \mathbb{P}(\tau) = IG(\tau|\alpha, \beta)$ and inverse-Wishart prior on the covariance matrix $\boldsymbol{S} \sim \mathbb{P}(\boldsymbol{S}) = IW(\boldsymbol{S}|\boldsymbol{\Psi}, v)$. If a sparse estimation of confounding effects is desired, selection or shrinkage priors can be assigned to $\boldsymbol{S}$, which we do not pursue in this paper.

Let $\mathcal{D} = \{(\boldsymbol{x}_i, z_i), i \in 1, \ldots, n\}$ be $n$ realizations of $(\boldsymbol{X}, Z)$. Denote $\boldsymbol{\beta} = [\beta_{j\ell k}]$ and $\boldsymbol{r} = [r_{j\ell}]$. The joint posterior distribution is then given by

$$\mathbb{P}(\boldsymbol{\beta}, \boldsymbol{S}, \boldsymbol{r}, \pi, \tau|\mathcal{D}) \propto \mathbb{P}(\boldsymbol{\beta}|\boldsymbol{r}, \tau)\mathbb{P}(\boldsymbol{S})\mathbb{P}(\boldsymbol{r}|\pi)\mathbb{P}(\pi)\mathbb{P}(\tau) \times \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{x}_i|z_i, \boldsymbol{B}(z_i), \boldsymbol{S}),$$

where $\mathbb{P}(\boldsymbol{\beta}|\boldsymbol{r}, \tau) = \prod_{j,\ell} \mathbb{P}(\boldsymbol{\beta}_{j\ell}|r_{j\ell}, \tau)$ and $\mathbb{P}(\boldsymbol{r}|\pi) = \prod_{j,\ell} \mathbb{P}(r_{j\ell}|\pi)$. The posterior distribution is not analytically available; we use Markov chain Monte Carlo (MCMC) to approximate it with the sampling steps detailed in Section S2 of Supplementary Materials.

The per-iteration computational complexity of sampling is $O(np^5)$[2]. This is a general learning algorithm that includes possible cycles and confounders. It can be simplified if there are no cycles and/or confounders. For example, for acyclic graphs, the spline coefficients $\boldsymbol{\beta}$ can be integrated out to improve MCMC mixing. Upon the completion of MCMC, the causal structure can be summarized by thresholding the estimated posterior probability of inclusion $\mathbb{P}(r_{j\ell} = 1|\mathcal{D}) \approx 1/M \sum_{m=1}^{M} I(r_{j\ell}^{(m)} = 1)$ at 0.5, where the superscript $(m)$ indexes the Monte Carlo samples.

## 4 EXPERIMENTS

We use extensive simulations as well as a real cancer genomic dataset with known cyclic causal graphs to evaluate the proposed method, CHOD. Additionally, we also applied CHOD to a Hong Kong stock market dataset analyzed in a recent heterogeneous causal discovery work [Huang et al., 2020]. Throughout, we set the hyperparameters as non-informative ones with $\boldsymbol{\Psi} = \boldsymbol{I}$, $v = p$, $a = b = 0.5$, $\alpha = \beta = 0.01$, and $K = 10$, which performed well in all experiments considered. We ran MCMC for 2000 iterations, discarded the first 1000 iterations as burn-in, and retained every 5th iteration after burn-in as posterior samples. We evaluated the graph structure recovery accuracy by calculating true positive rate (TPR), false discovery rate (FDR), and Matthew's correlation coefficient (MCC) based on 50 repetitions in simulations. MCC ranges over $[-1, 1]$ with 1 indicating perfect graph recovery.

### 4.1 SIMULATIONS

Since most existing causal discovery methods assume acyclic graphs and/or causal sufficiency with some exceptions that allow either cycles or confounders but usually not both, in order to maximize the fairness of comparison, we conducted simulations under three scenarios: when the simulation truths are cyclic graphs with confounders, acyclic graphs with confounders, and cyclic graphs without confounders, respectively. The first scenario is the most general one, which has been the focus of this paper, while the second and third scenarios are designed for fairness and hence are briefly discussed in the main text with details provided in the Section S3.1 of the Supplementary Materials. Note that our general algorithm accommodates all those three settings. We first considered data generated from our proposed model and

[2]Sampling $\boldsymbol{r}$ and $\boldsymbol{\beta}$ requires $O(p^2)$ numbers of likelihood evaluation and each likelihood evaluation is $O(np^3)$.

then considered various model misspecifications in terms of non-Gaussian errors, different confounding effects, varying degrees of heterogeneity, and unobserved covariates.

**Data generating mechanism** We considered sample size $n \in \{125, 250, 500, 1000\}$ and the number of nodes $p \in \{10, 25, 50\}$. Exogenous covariates were simulated from the uniform distribution $U(-1, 1)$. True causal graphs were generated as Erdős-Rényi random graph with edge probability $1/p$ (plotted in Figure S2-S4 in Section S3 of the Supplementary Materials). When assumed acyclic in the second scenario, the graph is constrained to have no directed cycles. Given the true structure, non-zero direct causal effects were randomly chosen from $f(Z) = 0.8Z$, $g(Z) = 0.9\cos(\pi Z)$, or $h(Z) = 0.9\tanh(\pi Z)$. We set the diagonal elements of $\boldsymbol{S}$ to 1. We generated the off-diagonal entries of $\boldsymbol{S}$ randomly from $U(-1, 1)$ in scenarios where there are unmeasured confounders, subject to $\boldsymbol{S}$ being positive-definite. Observations were then generated from model (1).

**Scenario 1: cyclic graphs with confounders** To the best of our knowledge, methods that can deal with both cycles and confounders in purely observational data are uncommon. We compared CHOD with two state-of-the-art acyclic causal discovery methods with confounders: RFCI [Colombo et al., 2012] and RICA [Salehkaleybar et al., 2020], and two state-of-the-art cyclic causal discovery methods without confounders: LiNG [Lacerda et al., 2008] and ANM [Mooij et al., 2011]. RICA and LiNG are based on linear non-Gaussian models, while ANM uses nonlinear additive noise models. RFCI imposes no distributional assumptions and outputs a graph containing both directed and bidirected edges (or edges with indeterminate directions). The results are summarized in Table 1. As expected, CHOD was the only approach that could recover the true graph well under this general heterogeneous simulation setting where both cycles and confounders are present. For example, the MCC for all the competing methods was uniformly low for all $(n, p)$ whereas the MCC of the proposed CHOD was always substantially higher and improved as sample size increased as expected.

**Scenario 2: acyclic graphs with confounders** In addition to RICA and RFCI, in this scenario, we compared CHOD with CAM [Bühlmann et al., 2014], GDS [Peters et al., 2014], and RESIT [Peters et al., 2014] as benchmarks although they are not designed for causal discovery in the presence of confounders. These three methods are based on nonlinear additive noise model. Moreover, we combined several bivariate causal discovery methods with CAM as suggested in Tagasovska et al. [2020] by first using CAM to learn a Markov equivalence class and then using IGCI [Janzing and Schölkopf, 2010], EMD [Chen et al., 2014], and bQCD [Tagasovska et al., 2020] to orient edges. These three bivariate causal discovery methods are based on asymmetry between the cause and the effect in terms of certain complexity metrics. In addition, we also compared with NOTEARS [Zheng et al., 2018] and DAG-GNN [Yu et al., 2019], which utilize continuous optimization for directed acyclic graph learning. Results are summarized in Table S1 in Section S3.1 of the Supplementary Materials. In summary, CHOD outperformed all the competing methods: the MCC of CHOD ranged from 0.6 to 0.9 whereas the competing methods had MCC typically <0.4. Furthermore, in Section S3.1 of the Supplementary Materials, we considered more comparisons with methods that incorporate the covariate $Z$ as an additional node in the causal graph (similar in spirit to the JCI framework [Mooij et al., 2020]). However, these additional comparisons did not show significantly better graph recovery.

**Scenario 3: cyclic graphs without confounders** We compared CHOD with LiNG and ANM. The results are summarized in Table S2 in Section S3.1 of the Supplementary Materials. As in the first two scenarios, CHOD performed significantly better by exploiting the data heterogeneity.

**Misspecification 1: nonlinear confounding, non-Gaussianity, and varying degrees of heterogeneity** From previous experiments, the proposed CHOD consistently outperformed non-Gaussian SEMs because data were heterogeneous and the errors were Gaussian, both conditions favoring CHOD. For fairer comparison and better illustration, we conducted further simulations under an alternative data generating mechanism. Specifically, we mimicked the simulation setting in Salehkaleybar et al. [2020] by generating $n = 250$ observations from the SEM (1) with uniform noises $e \sim U(-1, 1)$ under a three-node (Figure S5(b)) and a four-node (Figure S5(c)) directed acyclic graph, both having one unmeasured confounder (red, discarded at the model fitting stage). The non-zero direct causal effects were assumed to be quadratic with varying degrees of curvature (Figure S5(a)). As the degree of curvature approached zero, the data became more homogeneous. The coefficient functions were scaled to have the same average effects, that is we kept $\int |b(Z)| dZ$ constant.

We compared CHOD with RICA. Their receiver operating characteristic (ROC) curves under each true graph are shown in Figure S5(d) and S5(e) in the Supplementary Materials, respectively. In both cases, CHOD's performance deteriorated as the degree of heterogeneity *decreased*; by contrast, RICA's performance deteriorated as the degree of heterogeneity *increased*. Not surprisingly, when the data were completely homogeneous, CHOD was no better than a random guess in the three-node graph. However, somewhat surprisingly, in the four-node graph, CHOD still performed reasonably well even when the data were homogeneous. For instance, the worst area under the ROC curve were 0.865 and 0.789 for CHOD and RICA, respectively. Also note that, given sufficient degrees of heterogeneity, CHOD performed

Table 1: Simulation Scenario 1. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses. The best performance is shown in boldface.

| $n = 125$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.662 (0.065) | **0.253 (0.092)** | **0.644 (0.077)** | 0.662 (0.062) | **0.385 (0.080)** | **0.590 (0.094)** | 0.608 (0.082) | **0.385 (0.077)** | **0.590 (0.061)** |
| LiNG | **0.913 (0.088)** | 0.863 (0.012) | 0.104 (0.073) | **0.860 (0.057)** | 0.953 (0.003) | 0.023 (0.034) | **0.802 (0.059)** | 0.979 (0.002) | 0.007 (0.021) |
| ANM | 0.093 (0.069) | 0.879 (0.091) | 0.004 (0.083) | 0.002 (0.006) | 0.988 (0.035) | 0.009 (0.015) | 0.002 (0.006) | 0.972 (0.043) | 0.004 (0.022) |
| RFCI | 0.174 (0.069) | 0.742 (0.082) | 0.113 (0.071) | 0.227 (0.035) | 0.715 (0.046) | 0.200 (0.041) | 0.046 (0.025) | 0.943 (0.026) | 0.030 (0.025) |
| RICA | 0.470 (0.130) | 0.895 (0.029) | 0.051 (0.009) | 0.566 (0.104) | 0.947 (0.009) | 0.037 (0.045) | 0.485 (0.044) | 0.978 (0.002) | 0.006 (0.013) |
| $n = 250$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.804 (0.081) | **0.162 (0.088)** | **0.768 (0.075)** | 0.698 (0.065) | **0.286 (0.090)** | **0.662 (0.056)** | 0.680 (0.083) | **0.340 (0.081)** | **0.644 (0.086)** |
| LiNG | **0.880 (0.084)** | 0.870 (0.010) | 0.064 (0.068) | **0.834 (0.099)** | 0.955 (0.005) | 0.007 (0.051) | **0.798 (0.047)** | 0.979 (0.001) | 0.004 (0.017) |
| ANM | 0.077 (0.056) | 0.910 (0.057) | 0.024 (0.058) | 0.009 (0.013) | 0.937 (0.048) | 0.010 (0.033) | 0.003 (0.029) | 0.964 (0.012) | 0.004 (0.002) |
| RFCI | 0.201 (0.050) | 0.739 (0.045) | 0.123 (0.047) | 0.280 (0.033) | 0.744 (0.027) | 0.203 (0.030) | 0.009 (0.030) | 0.907 (0.027) | 0.069 (0.028) |
| RICA | 0.463 (0.127) | 0.890 (0.033) | 0.033 (0.099) | 0.481 (0.120) | 0.955 (0.011) | 0.001 (0.051) | 0.472 (0.065) | 0.978 (0.003) | 0.008 (0.019) |
| $n = 500$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.849 (0.073) | **0.152 (0.073)** | **0.813 (0.073)** | 0.813 (0.063) | **0.268 (0.090)** | **0.768 (0.094)** | **0.804 (0.075)** | **0.259 (0.057)** | **0.768 (0.061)** |
| LiNG | **0.917 (0.099)** | 0.867 (0.013) | 0.093 (0.083) | **0.875 (0.073)** | 0.952 (0.004) | 0.034 (0.039) | 0.788 (0.063) | 0.979 (0.002) | 0.002 (0.022) |
| ANM | 0.100 (0.066) | 0.898 (0.071) | 0.026 (0.074) | 0.009 (0.013) | 0.917 (0.058) | 0.009 (0.041) | 0.004 (0.008) | 0.961 (0.015) | 0.003 (0.018) |
| RFCI | 0.274 (0.045) | 0.734 (0.050) | 0.147 (0.052) | 0.320 (0.043) | 0.772 (0.030) | 0.197 (0.039) | 0.124 (0.027) | 0.901 (0.018) | 0.086 (0.022) |
| RICA | 0.413 (0.126) | 0.903 (0.033) | 0.078 (0.104) | 0.491 (0.094) | 0.954 (0.009) | 0.006 (0.040) | 0.489 (0.082) | 0.977 (0.004) | 0.015 (0.025) |
| $n = 1000$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | **0.912 (0.081)** | **0.142 (0.069)** | **0.840 (0.083)** | **0.894 (0.072)** | **0.245 (0.064)** | **0.813 (0.062)** | **0.849 (0.093)** | **0.251 (0.062)** | **0.786 (0.079)** |
| LiNG | 0.897 (0.108) | 0.866 (0.014) | 0.088 (0.089) | 0.836 (0.058) | 0.954 (0.003) | 0.013 (0.030) | 0.814 (0.058) | 0.978 (0.001) | 0.011 (0.020) |
| ANM | 0.127 (0.059) | 0.908 (0.042) | 0.037 (0.055) | 0.010 (0.006) | 0.903 (0.006) | 0.024 (0.006) | 0.003 (0.016) | 0.962 (0.017) | 0.009 (0.017) |
| RFCI | 0.303 (0.043) | 0.754 (0.034) | 0.139 (0.042) | 0.350 (0.028) | 0.792 (0.017) | 0.191 (0.022) | 0.159 (0.022) | 0.897 (0.013) | 0.101 (0.017) |
| RICA | 0.447 (0.142) | 0.893 (0.037) | 0.044 (0.114) | 0.506 (0.073) | 0.951 (0.007) | 0.018 (0.031) | 0.483 (0.078) | 0.977 (0.004) | 0.014 (0.023) |

reasonably well despite the non-Gaussianity and nonlinear confounding. For further comparison, we reran the analyses on data generated with Gaussian noises (keeping everything else the same). The results are shown in S5(g) and S5(e) in the Supplementary Materials. RICA were close to random guesses in both graphs whereas CHOD thrived on heterogeneity, especially in the four-node graph.

**Misspecification 2: partially homogeneous data and unknown covariates** We considered another type of model misspecification where data were partially homogeneous (i.e., observations were clustered and iid within each cluster) and the unknown covariate was estimated by UMAP. CHOD significantly outperformed the competing methods (see Section S3.2 of the Supplementary Materials).

## 4.2 APPLICATIONS

**Breast Cancer Genomic Data** We demonstrate the capability of CHOD in identifying gene feedback loops using breast cancer gene expression data downloaded from the Cancer Genome Atlas (`https://www.cancer.gov/tcga`). Breast cancer is a well-known extremely heterogeneous genetic disease. The dataset contains $n = 1215$ observations with 113 normal and 1102 tumor tissues. We focused on 8 feedback loops involving gene TP53 [Harris and Levine, 2005, plotted in Figure S7 in Section S3.3 of the Supplementary Materials]. We compared CHOD with two cyclic causal discovery methods, LiNG and ANM. In addition, we also considered two versions of JCI Mooij et al. [2020], FCI-JCI and ASD-JCI, which are broadly applicable for causal discovery with heterogeneous data. Gene expressions were log-transformed. For CHOD and JCI, we learned a one-dimensional embedding using UMAP as an input covariate and regressed out the effects of the covariate on the mean gene expression. The results are reported in Table 2. CHOD had uniformly strong performance: it was only outperformed by LiNG in one case in terms of MCC. See Section S3.3 of the Supplementary Materials for more elaborated description of the comparison.

**Hong Kong Stock Market dataset** We also apply CHOD to the Hong Kong stock market dataset analyzed by a recent heterogeneous causal discovery paper [Huang et al., 2020], containing 10 major stocks: HSBC Holdings plc (3), Hang Seng Bank Ltd (5), and Bank of China Hong Kong (Holdings) Ltd (10) from Hang Seng Finance Sub-index (HSF); Hong Kong Electric Holdings Limited (4) from Hang Seng Utilities Sub-index (HSU); Cheung Kong Holdings (1), Swire Group (8), and Cathay Pacific Airways Ltd (9) from Hang Seng Commerce & Industry Sub-index (HSC); Wharf (Holdings) Limited (2), Hong Kong Electric

Table 2: Application to breast cancer genomic data. 8 feedback loops involving gene TP53 were considered. The best performance is shown in boldface.

| Method | Network A | | | Network B | | | Network C | | | Network D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | **0.750** | 0.250 | 0.550 | 0.333 | **0.200** | **0.398** | **0.667** | **0.500** | **0.316** | **0.500** | 0.571 | **0.275** |
| LiNG | **0.750** | **0.000** | **0.791** | 0.583 | 0.563 | 0.198 | 0.333 | 0.667 | 0.000 | **0.500** | 0.667 | 0.164 |
| ANM | 0.500 | 0.333 | 0.316 | **0.667** | 0.556 | 0.238 | 0.333 | **0.500** | 0.189 | 0.167 | **0.500** | 0.180 |
| FCI-JCI | 0.250 | 0.500 | 0.059 | 0.417 | 0.500 | 0.219 | **0.667** | **0.500** | **0.316** | 0.333 | 0.667 | 0.123 |
| ASD-JCI | 0.500 | 0.333 | 0.316 | 0.500 | 0.455 | 0.298 | **0.667** | **0.500** | **0.316** | 0.167 | 0.750 | 0.010 |
| Method | Network E | | | Network F | | | Network G | | | Network H | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | **0.833** | **0.286** | **0.693** | 0.600 | **0.250** | **0.545** | **1.000** | **0.000** | **1.000** | **1.000** | **0.000** | **1.000** |
| LiNG | 0.333 | 0.778 | 0.031 | 0.800 | 0.500 | 0.405 | **1.000** | **0.000** | **1.000** | 0.500 | 0.000 | 0.577 |
| ANM | 0.667 | 0.556 | 0.359 | **1.000** | 0.583 | 0.389 | **1.000** | **0.000** | **1.000** | 0.500 | 0.000 | 0.577 |
| FCI-JCI | 0.500 | 0.700 | 0.114 | 0.400 | 0.667 | 0.035 | 0.000 | - | - | 0.000 | - | - |
| ASD-JCI | 0.500 | 0.625 | 0.217 | 0.600 | 0.571 | 0.221 | 0.000 | - | - | 0.000 | - | - |

Holdings Limited (4), and Sun Hung Kai Properties Limited (7) from Hang Seng Properties Sub-index (HSP). The result is shown in Figure 3, where the causal edges are consistent with some background market knowledge. As indicated by Huang et al. [2020], the within sub-index causal directions $5 \to 3$, $9 \to 8$ and $4 \to 1$ tend to follow the owner-member relationship. In addition, the following findings also match those in Huang et al. [2020]: stocks in HSF are major causes for those in HSC and HSP, and the stocks in HSP and HSU are major causes for those in HSC.

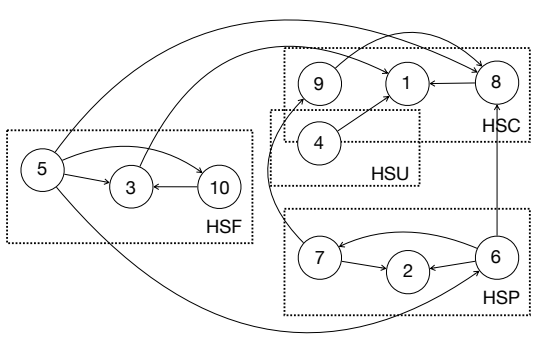



Figure 3: Application to the HK stock market dataset.

## 5 DISCUSSION

We have developed one of the first model-based causal discovery methods for observational data in the presence of both cyclic causality and confounders by exploiting the heterogeneity of causal mechanism. We have established several identifiability theories and carried out extensive experiments to demonstrate the utility of the proposed method against state-of-the-art alternatives.

There are many additional future directions can be taken to extend this work. For example, we have focused on linear Gaussian models, which enables efficient computation and causal effect estimation, and allows us to hopefully have conveyed the main idea that heterogeneity alone is enough to enable causal identification. Since nonlinearity and non-Gaussianity have already been proved useful for causal identification, we might have somewhat masked the contribution of heterogeneity if we had incorporated nonlinearity and/or non-Gaussianity into the proposed model. And as demonstrated in the experiments, the proposed model is relatively robust to non-Gaussian noises. That said, a natural future direction is indeed to extend this paper to nonlinear and non-Gaussian models via e.g., basis expansion and mixture of Gaussian error distributions.

Although we have not proven the most general case with causally insufficient multivariate cyclic graphs, our bivariate result is a strong indicator of identifiability in the general multivariate case because in the multivariate case, the existence of special graph structures (e.g., v-structure) can help causal identification whereas in the bivariate case, we can only rely on cause-effect asymmetry. We plan to prove the general identifiability in the future. Practically, as was done in many previous works like Tagasovska et al. [2020], one may first fix the skeleton or learn some partial structures with common structure learning algorithms, and then orient indeterminate edges by applying the bivariate causal discovery method to identify the full structure.

## References


Onn Brandman and Tobias Meyer. Feedback loops shape cellular signals in space and time. *Science*, 322(5900): 390–395, 2008.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Zhitang Chen and Laiwan Chan. Causal discovery for linear non-Gaussian acyclic models in the presence of latent Gaussian confounders. In *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation*, pages 17–24, 2012.

Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schölkopf. Causal discovery via reproducing kernel Hilbert space embeddings. *Neural Computation*, 26(7): 1484–1517, 2014.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

Patrick Forré and Joris M Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*, 2018.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in Neural Information Processing Systems*, 31:6266, 2018.

Sandra L Harris and Arnold J Levine. The p53 pathway: Positive and negative feedback loops. *Oncogene*, 24(17): 2899–2908, 2005.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 689–696, 2008a.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2): 362–378, 2008b.

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

Sui Huang et al. Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in Cell & Developmental Biology*, 20(7):869–876, 2009.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(109):3387–3439, 2012.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 366–374, 2008.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint 03426*, 2018.

Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 639–647, 2011.

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

Aristidis Moustakas and Carl-Henrik Heldin. Signaling networks guiding epithelial – mesenchymal transitions during embryogenesis and cancer progression. *Cancer Science*, 98(10):1512–1520, 2007.

Yang Ni, Yuan Ji, and Peter Müller. Reciprocal graphical models for integrative gene regulatory network analysis. *Bayesian Analysis*, 13(4):1095–1110, 2018.

Yang Ni, Francesco C Stingo, and Veerabhadran Baladandayuthapani. Bayesian graphical regression. *Journal of American Statistical Association*, 114(525):184–197, 2019.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15 (1):2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence*, pages 454–461, 1996.

Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. Causal structure discovery from distributions arising from mixtures of DAGs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8336–8345, 2020.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.

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

Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *Journal of Machine Learning Research*, 15:2629–2652, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9311–9323, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

Masanao Yajima, Donatello Telesca, Yuan Ji, and Peter Müller. Detecting differential patterns of interaction in molecular pathways. *Biostatistics*, 16(2):240–251, 2015.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

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

# Causal Discovery with Heterogeneous Observational Data: Supplementary Materials

## 1 PROOFS

Let $[n] = (1, \ldots, n)$. We call $ds(j) = \{\ell : \ell \leftrightarrow \cdots \leftrightarrow j\}$ the *districts* of $j$.

### 1.1 PROOF OF THEOREM 1

We prove it by contradiction. Suppose that $\mathcal{G} \neq \mathcal{G}'$ but

$$\mathbb{P}(\boldsymbol{X}|Z, \boldsymbol{B}(Z), \boldsymbol{S}) = \mathbb{P}(\boldsymbol{X}|Z, \boldsymbol{B}'(Z), \boldsymbol{S}'),\ \forall \boldsymbol{X}, Z,$$

for some $\boldsymbol{B}(Z), \boldsymbol{S}, \boldsymbol{B}'(Z), \boldsymbol{S}'$. Since centered Gaussian distribution is fully determined by its covariance, the two linear Gaussian SEMs are distribution equivalent if and only if

$$(\boldsymbol{I} - \boldsymbol{B}(Z))^T \boldsymbol{S}^{-1} (\boldsymbol{I} - \boldsymbol{B}(Z)) = (\boldsymbol{I} - \boldsymbol{B}'(Z))^T \boldsymbol{S}^{'-1} (\boldsymbol{I} - \boldsymbol{B}'(Z)),\ \forall Z,$$

which, in the bivariate case, is equivalent to the following three equations,

$$\begin{aligned}
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{11} b_{21}^2(Z) + 2\sigma_{12} b_{21}(Z) + \sigma_{22}) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{11} b'^2_{21}(Z) + 2\sigma'_{12} b'_{21}(Z) + \sigma'_{22}),\\
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{22} b_{12}^2(Z) + 2\sigma_{12} b_{12}(Z) + \sigma_{11}) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{22} b'^2_{12}(Z) + 2\sigma'_{12} b'_{12}(Z) + \sigma'_{11}),\\
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{11} b_{21}(Z) + \sigma_{22} b_{12}(Z) + \sigma_{12} b_{12}(Z) b_{21}(Z) + \sigma_{12})\\
&\qquad = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{11} b'_{21}(Z) + \sigma'_{22} b'_{12}(Z) + \sigma'_{12} b'_{12}(Z) b'_{21}(Z) + \sigma'_{12}).
\end{aligned} \tag{1}$$

If $\mathcal{G} \neq \mathcal{G}'$, then the equations above can at best have constant solutions, which contradicts our assumption. For example, since at least one but not all of $b_{12}(Z), b_{21}(Z), b'_{12}(Z), b'_{21}(Z)$ has to be zero because $\mathcal{G} \neq \mathcal{G}'$, without loss of generality, suppose $b_{12}(Z) = 0$ and $b'_{12}(Z) \neq 0$. Then the second equation of (1) is reduced to a quadratic equation of $b'_{12}(Z)$ of which the solutions are clearly constant in $Z$,

$$(\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{22} b'^2_{12}(Z) + 2\sigma'_{12} b'_{12}(Z) + \sigma'_{11}) - \sigma_{11}(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12}) = 0.$$

Therefore, $\mathcal{G} = \mathcal{G}'$. □

### 1.2 PROOF OF THEOREM 2

We first prove the identification of causal ordering by induction. Without loss of generality, we assume the true ordering is $[p]$. Letting initially the ordering $S = \emptyset$, we have

$$\mathrm{Var}(X_j|\boldsymbol{X}_S) = \mathrm{Var}(X_j) = \mathrm{Var}(\textstyle\sum_\ell b_{j\ell}(Z)X_\ell + \varepsilon_j),\ \forall j.$$

On the one hand, if $pa(j) = \emptyset$, i.e., $X_j$ is a root, then $\mathrm{Var}(X_j) = \mathrm{Var}(\varepsilon_j)$ is not a function of the exogenous covariate $Z$. On the other hand, if $pa(j) \neq \emptyset$, i.e., $X_j$ is not a root, $\mathrm{Var}(X_j)$ is a function of the covariate $Z$ by assumption. Hence we can pick a root node as the first of the causal ordering by examining whether $\mathrm{Var}(X_j)$ is a function of $Z$. Without loss of generality, we pick $X_1$.

Suppose we have picked the first $m$ nodes of the true ordering, $S = [m]$. Consider

$$\mathrm{Var}(X_j|\boldsymbol{X}_S) = \mathrm{Var}(\textstyle\sum_\ell b_{j\ell}(Z)X_\ell + \varepsilon_j|\boldsymbol{X}_S) = \mathrm{Var}(\textstyle\sum_{\ell>m} b_{j\ell}(Z)X_\ell + \varepsilon_j|\boldsymbol{X}_S),\ \forall j > m.$$

If $pa(j) \subseteq S$, i.e., $X_j$ is qualified as the next node of the causal ordering, then $\mathrm{Var}(X_j|\boldsymbol{X}_S) = \mathrm{Var}(\varepsilon_j|\boldsymbol{X}_S)$ is not a function of the covariate $Z$. By contrast, for any node that can not be the next in the ordering, $\mathrm{Var}(X_j|\boldsymbol{X}_S)$ is still a function of $Z$ by assumption. Hence we can identify the next node in the ordering by examining whether $\mathrm{Var}(X_j|\boldsymbol{X}_S)$ is a function of $Z$. Without loss of generality, we pick $j = m+1$ and set $S = [m+1]$ to be the first $m+1$ nodes of the correct ordering, which completes the proof of the ordering identifiability. Note that the causal ordering need not be unique but the constructive proof that we provide always identifies one such correct ordering.

Next, given the ordering $[p]$, we prove directed edges can be recovered if $pa(j) \cap ds(j) = \emptyset$. For the first node, we have $pa(1) = \emptyset$ and $\epsilon_1 = X_1$. For the second node, we have

$$\mathrm{Cov}(X_1, X_2) = b_{21}(Z)\mathrm{Var}(\epsilon_1) + \mathrm{Cov}(\epsilon_1, \epsilon_2).$$

If $pa(2) = \emptyset$, $\epsilon_2 = X_2$, and $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(\epsilon_1, \epsilon_2)$ is a not a function of the covariate $Z$. Otherwise, $pa(2) = \{1\}$, and we calculate $\epsilon_2 = X_2 - b_{21}(Z)X_1 = X_2 - \mathrm{Cov}(X_1, X_2)/\mathrm{Var}(\epsilon_1)X_1$, since $\mathrm{Cov}(\epsilon_1, \epsilon_2) = 0$ for $1 \notin ds(2)$.

Recursively, suppose we have identified the parent sets of the first $j-1$ nodes, the causal coefficients, and residuals. Denote $ds_{[j]}(j) = ds(j) \cap [j-1]$. Then for the $j$th node, when $\{ds_{[j]}(j) \cup pa(ds_{[j]}(j) \cup \{j\})\} \subseteq C \subseteq [j-1]$,

$$\mathrm{Cov}(X_k, X_j|\boldsymbol{X}_S) = \mathrm{Cov}(\textstyle\sum_{\ell \notin ds_{[j]}(j)} b_{\ell\to k}(Z)\epsilon_\ell + \epsilon_k, \epsilon_j|\boldsymbol{X}_C) = 0,\ \forall j > k \notin C, \tag{2}$$

where $b_{\ell\to k}(Z) = [(\boldsymbol{I} - \boldsymbol{B}(Z))^{-1}]_{k\ell}$ is the total causal effect from $X_\ell$ to $X_k$. Equivalently, when restricted to the first $j$ nodes, $\{ds_{[j]}(j) \cup pa(ds_{[j]}(j) \cup \{j\})\}$ is the Markov blanket of the $j$th node [Richardson, 2003]. We take the minimum set for which the conditional independence condition (2) is satisfied, then $C = \{ds_{[j]}(j) \cup pa(ds_{[j]}(j) \cap \{j\})\}$. For any $k \in C$,

$$\mathrm{Cov}(X_k, X_j|\boldsymbol{X}_{C\setminus\{k\}}) = \begin{cases} \mathrm{Cov}(\epsilon_k, \epsilon_j|\boldsymbol{X}_{C\setminus\{k\}}) = \mathrm{Cov}(\epsilon_k, \epsilon_j|\boldsymbol{\epsilon}_{ds_{[j]}(j)\setminus\{k\}}), & \text{if } k \in ds_{[j]}(j), \\ \mathrm{Cov}(X_k, b_{jk}(Z)X_k + \epsilon_j|\boldsymbol{X}_{C\setminus\{k\}}) & \\ \quad = b_{jk}(Z)\mathrm{Var}(X_k|\boldsymbol{X}_{pa(ds_{[j]}(j)\cup\{j\})\setminus\{k\}}), & \text{if } k \in pa(j). \end{cases}$$

The second quantity is a function of the covariate $Z$, whereas the first one is a constant. Therefore, we take the set $D = \{k : \mathrm{Cov}(X_k, X_j|\boldsymbol{X}_{C\setminus\{k\}}) = f(Z)\}$, then $pa(j) \subseteq D \subseteq C\setminus ds_{[j]}(j)$. Moreover,

$$\mathrm{Cov}(X_k, X_j|\boldsymbol{X}_{D\setminus\{k\}}, \boldsymbol{\epsilon}_{C\setminus D}) = \begin{cases} \mathrm{Cov}(X_k, \epsilon_j|\boldsymbol{X}_{D\setminus\{k\}}, \boldsymbol{\epsilon}_{C\setminus D}) = 0, & \text{if } k \in D\setminus pa(j), \\ \mathrm{Cov}(X_k, b_{jk}(Z)X_k + \epsilon_j|\boldsymbol{X}_{D\setminus\{k\}}, \boldsymbol{\epsilon}_{C\setminus D}) & \\ \quad = b_{jk}(Z)\mathrm{Var}(X_k|\boldsymbol{X}_{D\setminus\{k\}}, \boldsymbol{\epsilon}_{C\setminus D}) \neq 0, & \text{if } k \in pa(j). \end{cases}$$

We take $E = \{k : \mathrm{Cov}(X_k, X_j|\boldsymbol{X}_{D\setminus\{k\}}, \boldsymbol{\epsilon}_{C\setminus D}) \neq 0\}$, then $E = pa(j)$. Given the parent set, the causal coefficients and residuals can be easily computed, which completes the proof by induction. □

**Discussion of Theorem 2** Through direct calculation, we have for $S = [m], \forall m$,

$$\mathrm{Var}(X_j|\boldsymbol{X}_S) = \mathrm{Var}(\boldsymbol{A}_j\boldsymbol{\varepsilon}|\boldsymbol{\varepsilon}_S) = \boldsymbol{A}_{j,[p]\setminus S}(\boldsymbol{S}_{[p]\setminus S,[p]\setminus S} - \boldsymbol{S}_{[p]\setminus S,S}(\boldsymbol{S}_{S,S})^{-1}\boldsymbol{S}_{S,[p]\setminus S})\boldsymbol{A}^T_{j,[p]\setminus S},$$

where $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{B})^{-1}$ and $\boldsymbol{A}_j$ is the $j$th row of $\boldsymbol{A}$. By the definition of directed acyclic graphs, $A_{j\ell} \neq 0$ if and only if there exists a directed path from $X_\ell$ to $X_j$, i.e., $X_\ell$ is the ancestor of $X_j$. If $S$ contains all nodes precede $X_j$ in the causal ordering, $X_j$ is qualified as the next in the ordering and $\mathrm{Var}(X_j|\boldsymbol{X}_S)$ is not a function of $Z$. Otherwise, our assumption states that the covariate-dependent heterogeneous total causal effects from ancestors of $X_j$ in $[p]\backslash S$ to $X_j$ do not accidentally become homogeneous (i.e., the conditional variance is constant in $Z$). The variance dynamic allows us to identify the true causal ordering.

The additional assumption $pa(j) \cap ds(j) = \emptyset$ for causal graph identification is required to separate the heterogeneous effects from parents and districts (inherit from their patents). In fact, Maeda and Shimizu [2020] showed that their proposed method is only able to recover causal direction between pair of variables that are not affected by the same confounder. Wang and Drton [2020] proposed to learn causal graphs with unobserved confounders and non-Gaussian data, where the graphs are assumed to be simple acyclic mixed graphs. Our assumption is stronger but we believe it is due to the proof technique rather than the method itself which can be seen from good performance of CHOD in the simulations where the assumption $pa(j) \cap ds(j) = \emptyset$ was not enforced in generating the data or fitting the model. Theoretically relaxing this assumption will be our future work.

### 1.3 PROOF OF THEOREM 3

We say that $C \subseteq V$ is a cyclic component if it is a singleton or forms a directed cycle. A maximal cyclic component is a cyclic component such that none of its superset is a cyclic component. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of all maximal cyclic components. Since cycles are disjoint, it forms a partition of $V$. We define the collapsed graph $\widetilde{\mathcal{G}} = (\widetilde{V}, \widetilde{E})$ with $\widetilde{V} = \mathcal{C}$ (collapsing each maximal cyclic component to a single node) and $C_\ell \to C_j \in \widetilde{E}$ if and only if $c_r^\ell \to c_t^j$ for some $c_r^\ell \in C_\ell$ and $c_t^j \in C_j$. Then by construction, $\widetilde{\mathcal{G}}$ is acyclic. We assume without loss of generality that $(C_1, \ldots, C_k)$ is a topological ordering of $\widetilde{\mathcal{G}}$ and $c_1^\ell \to \cdots \to c_{|C_\ell|}^\ell \to c_1^\ell$ forms the maximal cyclic component $C_\ell$. Denote $C_\ell^+ = C_{\ell+1} \cup \cdots \cup C_k$. For any (ordered) sets $C = (c_\ell)$ and $D = (d_k) \subseteq V$, let $\boldsymbol{B}_{D,D}(Z)$ be the submatrix of $\boldsymbol{B}(Z)$ with rows and columns indexed by $D$, $\boldsymbol{A}_{D,D}(Z) = (\boldsymbol{I} - \boldsymbol{B}_{D,D}(Z))^{-1}$, $\boldsymbol{E}_{D,C} = (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_{|D|}^T)^T$ with $e_{k,\ell} = 1$ if $d_k = c_\ell$ and $e_{k,\ell} = 0$ otherwise.

We first constructively prove that the ordering of maximal cyclic components and the edge directions within each maximal cyclic component are identifiable. Suppose we have identified the first $\ell - 1$ maximal cyclic components for $\ell = 1, \ldots, k$, and we are looking for the next candidate $D \subseteq C_{\ell-1}^+ : d_1 \to \cdots \to d_{|D|} \to d_1$ in the ordering. Because of causal sufficiency, $D$ is a valid candidate (i.e., it complies with a true ordering and the edge direction in $D$ matches the truth) if there exists a transformation matrix

$$\boldsymbol{A}_{D,D}^{'-1}(Z) = \boldsymbol{I} - \boldsymbol{B}'_{D,D}(Z) = \begin{bmatrix} 1 & 0 & \cdots & 0 & -b'_{d_1,d_{|D|}}(Z) \\ -b'_{d_2,d_1}(Z) & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -b'_{d_{|D|},d_{|D|-1}}(Z) & 1 \end{bmatrix}$$

such that $\mathrm{Cov}(\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}) = \mathrm{diag}(\sigma'_{d_1}, \ldots, \sigma'_{d_{|D|}})$. Therefore, we formulate the following condition:

**Condition** $(\star)$: for any $D$ that cannot be the next maximal cyclic component in the ordering, there does not exist a transformation $\boldsymbol{A}_{D,D}^{'-1}(Z)$ such that $\mathrm{Cov}(\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}})$ is a diagonal matrix.

Notice that when $D = C_\ell$ which is a validate candidate, we can choose $\boldsymbol{A}_{D,D}^{'-1}(Z) = \boldsymbol{A}_{C_\ell,C_\ell}^{-1}(Z)$ which leads to $\mathrm{Cov}(\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}) = \mathrm{Cov}(\boldsymbol{\epsilon}_{C_\ell}) = \mathrm{diag}(\sigma_{c_1^\ell}, \ldots, \sigma_{c_{|C_\ell|}^\ell})$; hence $C_\ell$ is a valid next maximal cyclic component.

For any set $D = (d_1, \ldots, d_{|D|}) \subseteq C_{\ell-1}^+$, we have

$$\begin{aligned} \boldsymbol{X}_{D\cap C_\ell} &= \boldsymbol{E}_{D\cap C_\ell,C_\ell}\boldsymbol{A}_{C_\ell,C_\ell}(Z)\boldsymbol{\epsilon}_{C_\ell} + \boldsymbol{F}(\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}), \\ \boldsymbol{X}_{D\cap C_\ell^+} &= \boldsymbol{E}_{D\cap C_\ell^+,C_\ell^+}\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)[\boldsymbol{B}_{C_\ell^+,C_\ell}(Z)\boldsymbol{A}_{C_\ell,C_\ell}(Z)\boldsymbol{\epsilon}_{C_\ell} + \boldsymbol{\epsilon}_{C_\ell^+}] + \boldsymbol{F}(\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}), \end{aligned}$$

where $\boldsymbol{\epsilon}_{C_{\ell-1}^+} \perp \boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}$ and $\boldsymbol{F}(\cdot)$ is some deterministic function which will become zero when taking the

conditional covariance later on (its complex functional form is irrelevant here and hence not shown). Therefore,

$$
\begin{aligned}
&[\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D]_k = m(\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}}) + \\
&\begin{cases}
[\boldsymbol{E}_{d_k,C_\ell} - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell}]\boldsymbol{A}_{C_\ell,C_\ell}(Z)\boldsymbol{\epsilon}_{C_\ell}, & \text{if } d_{k-1}\in C_\ell, d_k\in C_\ell, \\
[\boldsymbol{E}_{d_k,C_\ell} - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell^+}\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{B}_{C_\ell^+,C_\ell}(Z)]\boldsymbol{A}_{C_\ell,C_\ell}(Z) & \\
\qquad \times\boldsymbol{\epsilon}_{C_\ell} - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell^+}\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{\epsilon}_{C_\ell^+}, & \text{if } d_{k-1}\in C_\ell^+, d_k\in C_\ell, \\
[\boldsymbol{E}_{d_k,C_\ell^+}\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{B}_{C_\ell^+,C_\ell}(Z) - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell}]\boldsymbol{A}_{C_\ell,C_\ell}(Z) & \\
\qquad \times\boldsymbol{\epsilon}_{C_\ell} + \boldsymbol{E}_{d_k,C_\ell^+}\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{\epsilon}_{C_\ell^+}, & \text{if } d_{k-1}\in C_\ell, d_k\in C_\ell^+, \\
[\boldsymbol{E}_{d_k,C_\ell^+} - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell^+}]\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{B}_{C_\ell^+,C_\ell}(Z)\boldsymbol{A}_{C_\ell,C_\ell}(Z) & \\
\qquad \times\boldsymbol{\epsilon}_{C_\ell} + [\boldsymbol{E}_{d_k,C_\ell^+} - b'_{d_k,d_{k-1}}(Z)\boldsymbol{E}_{d_{k-1},C_\ell^+}]\boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z)\boldsymbol{\epsilon}_{C_\ell^+}, & \text{if } d_{k-1}\in C_\ell^+, d_k\in C_\ell^+,
\end{cases}
\end{aligned}
$$

where $m(\cdot)$ is some deterministic function (its functional form is omitted for the same reason as above). Eliminating $b'_{d_k,d_{k-1}}(Z)$ from the set of equations induced by $\mathrm{Cov}(\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}}) = \mathrm{diag}(\sigma'_{d_1},\ldots,\sigma'_{d_{|D|}})$ for any invalid candidate $D$ introduces a peculiar constraint on the causal effect functions:

$$f(\boldsymbol{A}_{C_\ell,C_\ell}(Z), \boldsymbol{A}_{C_\ell^+,C_\ell^+}(Z), \boldsymbol{B}_{C_\ell^+,C_\ell}(Z)) = 0$$

for certain $f(\cdot)$. The condition $(\star)$ then rules out such peculiar situation.

Next, given the ordering of the maximal cyclic components, we have

$$\mathrm{Cov}(\boldsymbol{X}_{C_{\ell+1}}, \boldsymbol{X}_{C_\ell}|\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}}) = \boldsymbol{A}_{C_{\ell+1},C_{\ell+1}}\boldsymbol{B}_{C_{\ell+1},C_\ell}\mathrm{Var}(\boldsymbol{X}_{C_\ell}|\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}}).$$

Therefore,

$$\boldsymbol{B}_{C_{\ell+1},C_\ell} = \boldsymbol{A}_{C_{\ell+1},C_{\ell+1}}^{-1}\mathrm{Cov}(\boldsymbol{X}_{C_{\ell+1}}, \boldsymbol{X}_{C_\ell}|\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}})\mathrm{Var}^{-1}(\boldsymbol{X}_{C_\ell}|\boldsymbol{X}_{C_1},\ldots,\boldsymbol{X}_{C_{\ell-1}}),$$

and hence the directed edges from component $C_\ell$ to $C_{\ell+1}$ can be recovered from $\boldsymbol{B}_{C_{\ell+1},C_\ell}$. □

**Discussion of Theorem 3** We illustrate the condition $(\star)$ in the proof of Theorem 3 with a toy example. Consider the graph $\mathcal{G}$ in Figure 1. The maximal cyclic components are $C_1 = \{1,2,3\}$ and $C_2 = \{4,5\}$. The collapsed graph $\widetilde{\mathcal{G}}$ is simply $C_1 \to C_2$.

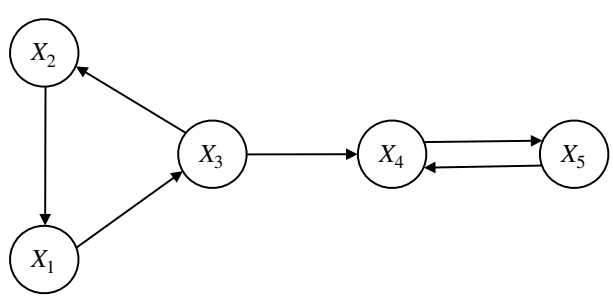



Figure 1: A demonstrative example of Theorem 3.

If $D = \{3\}$, we have

$$X_3 = [b_{31}(Z)\epsilon_1 + b_{31}(Z)b_{12}(Z)\epsilon_2 + \epsilon_3]/[1 - b_{12}(Z)b_{23}(Z)b_{31}(Z)].$$

Therefore, $\mathrm{Var}(X_3)$ is in general not constant in $Z$.

If $D = \{3,4\}$, we have

$$
\begin{aligned}
X_4 = \{&b_{43}(Z)[b_{31}(Z)\epsilon_1 + b_{31}(Z)b_{12}(Z)\epsilon_2 + \epsilon_3]/[1 - b_{12}(Z)b_{23}(Z)b_{31}(Z)] \\
&+ [\epsilon_4 + b_{45}(Z)\epsilon_5]\}/[1 - b_{45}(Z)b_{54}(Z)].
\end{aligned}
$$

The violation of condition ($\star$), i.e., there exists some $\boldsymbol{A}_{D,D}^{'-1}$ such $\mathrm{Cov}(\boldsymbol{A}_{D,D}^{'-1}(Z)\boldsymbol{X}_D)$ is a diagonal matrix, gives rise to the following three conditions:

$$\begin{aligned}
&\mathrm{Var}(X_3 - b'_{34}(Z)X_4) \text{ is constant in } Z,\\
&\mathrm{Var}(X_4 - b'_{43}(Z)X_3) \text{ is constant in } Z,\\
&\mathrm{Cov}(X_3 - b'_{34}(Z)X_4, X_4 - b'_{43}(Z)X_3) = 0.
\end{aligned}$$

which reduces to one condition after eliminating $b'_{34}(Z)$ and $b'_{43}(Z)$,

$$\begin{aligned}
0 &= \mathrm{Cov}(X_3, X_4) - \mathrm{Cov}(X_3, X_4)[\mathrm{Cov}(X_3, X_4) \pm (\mathrm{Cov}^2(X_3, X_4) - \mathrm{Var}(X_4)(\mathrm{Var}(X_3) - a))^{1/2}]\\
&\times [\mathrm{Cov}(X_3, X_4) \pm (\mathrm{Cov}^2(X_3, X_4) - \mathrm{Var}(X_3)(\mathrm{Var}(X_4) - b))^{1/2}]/[\mathrm{Var}(X_3)\mathrm{Var}(X_4)]\\
&\pm (\mathrm{Cov}^2(X_3, X_4) - \mathrm{Var}(X_4)(\mathrm{Var}(X_3) - a))^{1/2} \pm (\mathrm{Cov}^2(X_3, X_4) - \mathrm{Var}(X_3)(\mathrm{Var}(X_4) - b))^{1/2},
\end{aligned}$$

where $a, b$ are constants, and

$$\begin{aligned}
\mathrm{Var}(X_3) &= [b_{31}^2(Z)\sigma_1 + b_{31}^2(Z)b_{12}^2(Z)\sigma_2 + \sigma_3]/[1 - b_{12}(Z)b_{23}(Z)b_{31}(Z)]^2,\\
\mathrm{Var}(X_4) &= \{b_{43}^2(Z)[b_{31}^2(Z)\sigma_1 + b_{31}^2(Z)b_{12}^2(Z)\sigma_2 + \sigma_3]/[1 - b_{12}(Z)b_{23}(Z)b_{31}(Z)]^2\\
&\quad + \sigma_4 + b_{45}^2(Z)\sigma_5\}/[1 - b_{45}(Z)b_{54}(Z)]^2 = [b_{43}^2(Z)\mathrm{Var}(X_3) + \sigma_4 + b_{45}^2(Z)\sigma_5]/[1 - b_{45}(Z)b_{54}(Z)]^2,\\
\mathrm{Cov}(X_3, X_4) &= b_{43}(Z)[b_{31}^2(Z)\sigma_1 + b_{31}^2(Z)b_{12}^2(Z)\sigma_1 + \sigma_3]/\{[1 - b_{12}(Z)b_{23}(Z)b_{31}(Z)]^2\\
&\quad [1 - b_{45}(Z)b_{54}(Z)]\} = b_{43}(Z)\mathrm{Var}(X_3)/[1 - b_{45}(Z)b_{54}(Z)].
\end{aligned}$$

Hence, unless the covariate-dependent direct causal effects satisfy this peculiar equation, one will not mistakenly identify $D = \{3, 4\}$ as a valid maximal cyclic component.

If $D = \{1, 2, 3\}$, but the cycle direction is reversed: $1 \to 2 \to 3 \to 1$. Then condition ($\star$) implies that there do not exist constants $a, b, c, d$ such that

$$b_{12}(Z) = a \cdot b_{23}(Z) + b = c \cdot b_{31}(Z) + d.$$

Therefore, unless $b_{12}(Z), b_{23}(Z), b_{31}(Z)$ happen to be linear transformation of each, one will not mistakenly identify the reversed cycle direction.

### 1.4 PROOF OF PROPOSITION 1

According to our assumption, given the graph structure $\mathcal{G}$ the following transformations

$$\boldsymbol{\phi} : \boldsymbol{m}(\boldsymbol{Z}) \mapsto \mathbb{P}(\boldsymbol{X}|\boldsymbol{m}(\boldsymbol{Z}), \boldsymbol{S}),\ \boldsymbol{m} : \boldsymbol{Z} \mapsto \boldsymbol{m}(\boldsymbol{Z})$$

are continuous and injective. Therefore, the composite mapping $\boldsymbol{\psi} := \boldsymbol{\phi} \circ \boldsymbol{m} : \boldsymbol{Z} \mapsto \mathbb{P}(\boldsymbol{X}|\boldsymbol{m}(\boldsymbol{Z}), \boldsymbol{S})$ is continuous and injective, so do its univariate marginals. Then the monotonicity of $\boldsymbol{\psi}$ follows. □

## 2 MCMC ALGORITHM

The proposed MCMC algorithm repeats the following five steps until convergence.

1. We generate the covariance matrix $\boldsymbol{S}$ of noises from the full conditional distribution

$$\boldsymbol{S} \sim IW(\boldsymbol{\Psi}', v'),\ \boldsymbol{\Psi}' = \boldsymbol{\Psi} + \textstyle\sum_{i=1}^n \{\boldsymbol{x}_i - \boldsymbol{B}(z_i)\boldsymbol{x}_i\}\{\boldsymbol{x}_i - \boldsymbol{B}(z_i)\boldsymbol{x}_i\}^T,\ v' = v + n.$$

2. We sample each edge by a reversible jump (birth-death) step. For each $j \neq \ell = 1, \ldots, p$, we propose a new state $r'_{j\ell} = 1 - r_{j\ell}$. If $r'_{j\ell} = 0$ (death move), set $\boldsymbol{\beta}'_{j\ell} = \boldsymbol{0}$. Otherwise (birth move), sample $\boldsymbol{\beta}'_{j\ell} \sim N(\boldsymbol{0}, \tau\boldsymbol{I})$. Accept the new $(r'_{j\ell}, \boldsymbol{\beta}'_{j\ell})$ with probability $\min(\alpha, 1)$, where

$$\log\alpha = (-1)^{r'_{j\ell}} \log\tfrac{1-\pi}{\pi} + \textstyle\sum_{i=1}^n \big\{\log\mathbb{P}(\boldsymbol{x}_i|z_i, \boldsymbol{B}'(z_i), \boldsymbol{S}) - \mathbb{P}(\boldsymbol{x}_i|z_i, \boldsymbol{B}(z_i), \boldsymbol{S})\big\},$$

with $\boldsymbol{B}'(z) = \boldsymbol{B}(z)$ except for the entry being updated, $b'_{j\ell}(z) = \sum_{k=1}^K \beta'_{j\ell k}\phi_k(z)$.

3. We sample non-zero spline coefficients by a Metropolis-Hasting step. For each $j \neq \ell = 1, \ldots, p$ and $k = 1, \ldots, K$, we propose non-zero $\beta_{j\ell k}$ (corresponds to $r_{j\ell} = 1$) by a random walk proposal density centered at the current value $\beta'_{j\ell k} \sim N(\beta_{j\ell k}, \sigma)$. Accept the new $\beta'_{j\ell k}$ with probability $\min(\alpha, 1)$, where

$$\begin{aligned}\log \alpha = &\log \mathbb{P}(\beta'_{j\ell k} | r_{j\ell} = 1, \tau) - \log \mathbb{P}(\beta_{j\ell k} | r_{j\ell} = 1, \tau) \\ &+ \textstyle\sum_{i=1}^n \big\{ \log \mathbb{P}(\boldsymbol{x}_i | z_i, \boldsymbol{B}'(z_i), \boldsymbol{S}) - \mathbb{P}(\boldsymbol{x}_i | z_i, \boldsymbol{B}(z_i), \boldsymbol{S}) \big\},\end{aligned}$$

with $\boldsymbol{B}'(z) = \boldsymbol{B}(z)$ except for the entry currently being updated, $b'_{j\ell}(z) = \beta'_{j\ell k}\phi_k(z) + \sum_{h \neq k} \beta_{j\ell h}\phi_h(z)$.

4. We generate the variance of non-zero coefficients $\tau$ from the full conditional distribution

$$\tau \sim IG(\alpha', \beta'),\ \alpha' = \alpha + \tfrac{1}{2} \textstyle\sum_{j,\ell,k} I(\beta_{j\ell k} \neq 0),\ \beta' = \beta + \tfrac{1}{2} \sum_{j,\ell,k} \beta^2_{j\ell k}.$$

5. We generate the edge inclusion probability $\pi$ from the full conditional distribution

$$\pi \sim \text{beta}(a', b'),\ a' = a + \textstyle\sum_{j \neq \ell} r_{j\ell},\ b' = b + \sum_{j \neq \ell} (1 - r_{j\ell}).$$

### 2.1 IMPLEMENTATION OF CHOD WITH LATENT COVARIATES

Suppose $Z$ is univariate and latent. Our Bayesian formulation can be easily adapted for the joint estimation of $Z$ and causal graphs. Without loss of generality, we assume $Z \in [0, 1]$. We assign independent uniform prior $z_i \sim U(0, 1)$ or the Coulomb repulsive prior [Wang and Dunson, 2015] for better separation

$$\mathbb{P}(z_1, \ldots, z_n) \propto \textstyle\prod_{j=i+1}^n \sin^{2\gamma}\{\pi(z_i - z_j)\},\ \forall z_i \in [0, 1]$$

with the repulsive parameter $\gamma$. For MCMC implementation, we add the following step to sample $z_1, \ldots, z_n$ independently

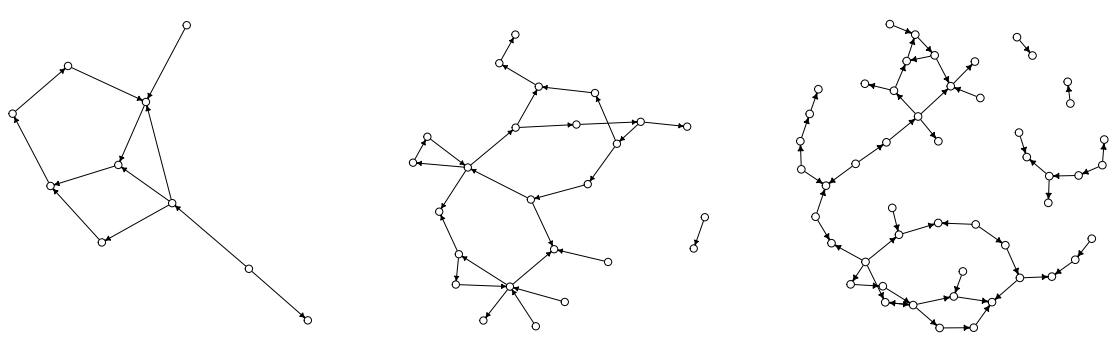
- We propose $z'_i \sim \mathbb{Q}(z'_i | z_i)$ and accept it with probability $\min(1, \alpha)$, where

$$\begin{aligned}\log \alpha = &\log\{\mathbb{Q}(z_i | z'_i)\mathbb{P}(z'_i, \boldsymbol{z}_{-i})\mathbb{P}(\boldsymbol{x}_i | z'_i, \boldsymbol{B}(z'_i), \boldsymbol{S})\} \\ &- \log\{\mathbb{Q}(z'_i | z_i)\mathbb{P}(z_i, \boldsymbol{z}_{-i})\mathbb{P}(\boldsymbol{x}_i | z_i, \boldsymbol{B}(z_i), \boldsymbol{S})\}.\end{aligned}$$

  In the above, $\mathbb{Q}(z'_i | z_i)$ is a random walk proposal density truncated at $[0, 1]$.

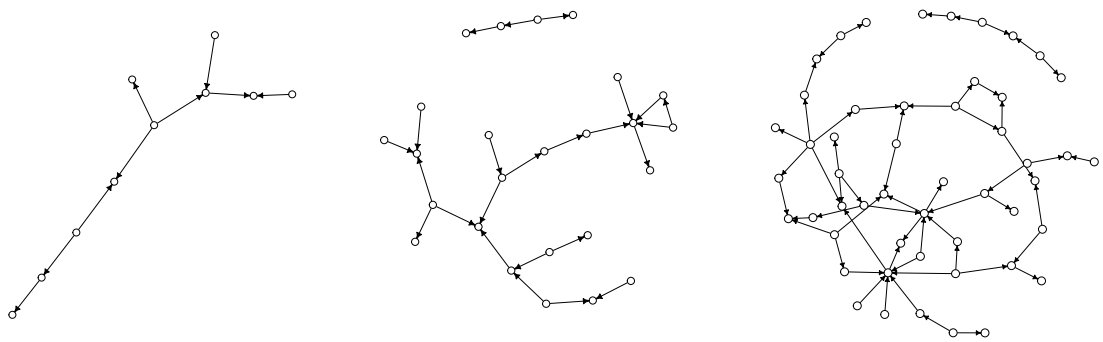
## 3 ADDITIONAL DETAILS OF THE EXPERIMENTS

Figures 2–4 show the randomly generated simulation true causal graphs in Scenarios 1–3.



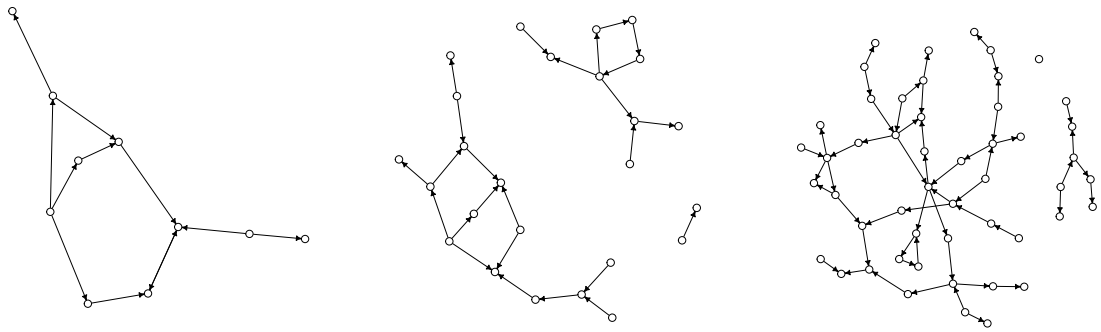
(a) $p = 10$. (b) $p = 25$. (c) $p = 50$.

Figure 2: Simulation true graphs in Scenario 1 (cyclic graphs with confounders).



(a) $p = 10$. (b) $p = 25$. (c) $p = 50$.

Figure 3: Simulation true graphs in Scenario 2 (acyclic graphs with confounders).



(a) $p = 10$. (b) $p = 25$. (c) $p = 50$.

Figure 4: Simulation true graphs in Scenario 3 (cyclic graphs without confounders).

### 3.1 ADDITIONAL DETAILS OF SCENARIO 2&3

Table 1 and 2 respectively show summaries for simulation Scenario 2 and 3. Clearly, CHOD outperformed others by a significant amount. Additionally, we used scenario 2 (acyclic graphs with confounders) with $p = 10$ and $n \in \{125, 500\}$ to illustrate the comparison with alternative methods (RFCI, RICA, CAM, GDS, RESIT, IGCI, EMD, and bQCD as in the main text), where $Z$ was included as a graph node. Results are shown in Table 3. The conclusion stays the same: CHOD outperforms the alternatives with larger TPR and smaller FDR.

### 3.2 ADDITIONAL DETAILS OF MODEL MISSPECIFICATION

**Misspecification 1** Figure 5 showed the result of model misspecification 1.

**Misspecification 2** We considered a dataset with $n = 250$ observations which were assigned to $K = 10$ clusters uniformly at random. We considered the two causal graphs previously used in model misspecification 1, of which the structures were assumed to be the same across clusters but the causal effects were different. Within each cluster, we generated $Z_k$ uniformly from $[-k/10, -(k-1)/10] \cup [(k-1)/10, k/10]$ and $\boldsymbol{X}_k$ from the following SEM for $k = 1, \ldots, K$,

$$\boldsymbol{X}_k = \boldsymbol{D}(Z_k) + \boldsymbol{B}_k \boldsymbol{X}_k + \boldsymbol{\mathcal{E}}_k, \;\; \boldsymbol{\mathcal{E}}_k \sim N(\boldsymbol{0}, \boldsymbol{I}),$$

where $\boldsymbol{D}(Z) = [d_j(Z)]$ and $\boldsymbol{B}_k = [b_{j\ell k}]$. We set $d_j(Z) = Z, \forall j$ and non-zero coefficients $b_{j\ell k} = k/10$. Confounders were again discarded at the model fitting stage and $Z$ was unobserved.

For CHOD, we first imputed $Z$ by UMAP. Then the mean effects of $Z$ were regressed out. We compared CHOD with RICA and CAM. The results are shown in Figure 6. Despite the fact the data were partially homogeneous, the confounding

effects were non-constant (vary across clusters), and exogenous covariates were unknown, CHOD combined with UMAP substantially outperformed the competing methods with AUC 0.980 and 0.957 for the three-node and the four-node graphs, respectively.

### 3.3 ADDITIONAL RESULTS FOR THE APPLICATION

The estimated networks from CHOD are shown in Figure 7. CHOD performed especially well on the PTEN/AKT/MDM-2 loop (Network E).

## 4 BRIEF DISCUSSION OF CAUSAL INFERENCE WITH CHOD

When making inference like finding post-intervention distribution $\mathbb{P}(\boldsymbol{Y}|do(\boldsymbol{W}))$ for $\boldsymbol{Y}, \boldsymbol{W} \subseteq \boldsymbol{X}$, linear Gaussianity allows analytical marginalization for causal inference, whereas complex models such as linear non-Gaussian and nonlinear Gaussian models do not (discrete approximation is often required which is #P-hard and sampling-based approximation is still NP-hard). For example, for an acyclic graph, the causal effect of $do(X_j = x_j)$ can be computed as [Maathuis et al., 2009]:

$$\frac{\partial}{\partial x}\mathbb{E}(X_k|do(X_j = x), Z)|_{x=x_j} = [\boldsymbol{\Sigma}(Z)_{k,pa^+(j)}\boldsymbol{\Sigma}(Z)^{-1}_{pa^+(j),pa^+(j)}]_1,$$

where $\boldsymbol{\Sigma}(Z)$ is the covariance matrix of $\boldsymbol{X}$ given $Z$ and $pa^+(j) = \{j\} \cup pa(j)$.

### References


Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Takashi Nicholas Maeda and Shohei Shimizu. RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 735–745, 2020.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1): 145–157, 2003.

Y Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *arXiv preprint 11131*, 2020.

Ye Wang and David B Dunson. Probabilistic curve learning: Coulomb repulsion and the electrostatic Gaussian process. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1738–1746, 2015.

Table 1: Simulation Scenario 2. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses. The best performance is shown in boldface.

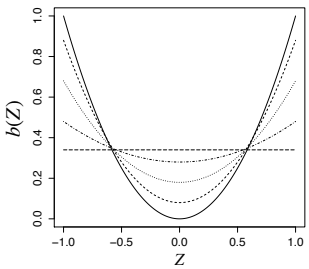
| $n = 125$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.659 (0.114) | **0.258 (0.103)** | **0.656 (0.080)** | 0.625 (0.076) | **0.243 (0.093)** | **0.644 (0.078)** | **0.541 (0.055)** | **0.274 (0.089)** | **0.572 (0.060)** |
| RFCI | 0.314 (0.097) | 0.393 (0.165) | 0.373 (0.127) | 0.288 (0.044) | 0.524 (0.063) | 0.332 (0.050) | 0.093 (0.031) | 0.849 (0.049) | 0.097 (0.039) |
| RICA | 0.569 (0.091) | 0.754 (0.041) | 0.223 (0.077) | 0.435 (0.053) | 0.889 (0.014) | 0.089 (0.033) | 0.436 (0.053) | 0.945 (0.010) | 0.080 (0.021) |
| CAM | 0.436 (0.157) | 0.899 (0.036) | 0.029 (0.101) | 0.323 (0.073) | 0.940 (0.014) | 0.058 (0.036) | 0.199 (0.037) | 0.959 (0.017) | 0.042 (0.018) |
| GDS | 0.214 (0.134) | 0.769 (0.139) | 0.148 (0.139) | 0.368 (0.119) | 0.706 (0.132) | 0.257 (0.087) | 0.335 (0.079) | 0.750 (0.042) | 0.255 (0.035) |
| RESIT | 0.058 (0.065) | 0.827 (0.199) | 0.057 (0.110) | 0.018 (0.033) | 0.941 (0.110) | 0.013 (0.060) | 0.058 (0.040) | 0.829 (0.131) | 0.085 (0.072) |
| IGCI | 0.107 (0.039) | 0.540 (0.138) | 0.188 (0.074) | 0.150 (0.084) | 0.570 (0.302) | 0.166 (0.092) | 0.063 (0.033) | 0.900 (0.051) | 0.062 (0.042) |
| EMD | 0.004 (0.022) | 0.980 (0.100) | 0.018 (0.051) | 0.108 (0.069) | 0.634 (0.339) | 0.113 (0.075) | 0.075 (0.036) | 0.881 (0.054) | 0.078 (0.044) |
| bQCD | 0.004 (0.022) | 0.990 (0.050) | 0.016 (0.074) | 0.058 (0.056) | 0.707 (0.381) | 0.050 (0.069) | 0.050 (0.021) | 0.920 (0.033) | 0.046 (0.027) |
| NOTEARS | 0.785 (0.033) | 0.819 (0.026) | 0.252 (0.028) | **0.839 (0.021)** | 0.933 (0.028) | 0.175 (0.024) | 0.267 (0.029) | 0.948 (0.043) | 0.045 (0.011) |
| DAG-GNN | **0.801 (0.029)** | 0.773 (0.021) | 0.311 (0.023) | 0.837 (0.041) | 0.824 (0.038) | 0.203 (0.035) | 0.462 (0.035) | 0.897 (0.048) | 0.175 (0.039) |
| $n = 250$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.718 (0.118) | **0.237 (0.094)** | **0.695 (0.086)** | 0.696 (0.081) | **0.205 (0.070)** | **0.705 (0.060)** | **0.659 (0.086)** | **0.216 (0.061)** | **0.713 (0.065)** |
| RFCI | 0.425 (0.081) | 0.407 (0.127) | 0.433 (0.091) | 0.369 (0.034) | 0.536 (0.051) | 0.372 (0.037) | 0.148 (0.030) | 0.836 (0.034) | 0.130 (0.033) |
| RICA | 0.574 (0.114) | 0.757 (0.046) | 0.209 (0.088) | 0.547 (0.070) | 0.879 (0.015) | 0.127 (0.041) | 0.565 (0.044) | 0.946 (0.004) | 0.093 (0.017) |
| CAM | 0.533 (0.246) | 0.893 (0.049) | 0.053 (0.156) | 0.351 (0.059) | 0.945 (0.013) | 0.059 (0.031) | 0.164 (0.051) | 0.960 (0.013) | 0.037 (0.027) |
| GDS | 0.243 (0.093) | 0.767 (0.087) | 0.159 (0.099) | 0.325 (0.134) | 0.713 (0.121) | 0.225 (0.066) | 0.286 (0.103) | 0.751 (0.082) | 0.214 (0.090) |
| RESIT | 0.129 (0.083) | 0.738 (0.162) | 0.128 (0.116) | 0.033 (0.059) | 0.869 (0.095) | 0.078 (0.075) | 0.092 (0.061) | 0.852 (0.097) | 0.076 (0.075) |
| IGCI | 0.111 (0.064) | 0.783 (0.117) | 0.094 (0.085) | 0.135 (0.039) | 0.847 (0.057) | 0.112 (0.051) | 0.099 (0.038) | 0.868 (0.047) | 0.097 (0.042) |
| EMD | 0.111 (0.091) | 0.784 (0.174) | 0.091 (0.131) | 0.167 (0.068) | 0.815 (0.075) | 0.144 (0.074) | 0.107 (0.033) | 0.857 (0.036) | 0.106 (0.034) |
| bQCD | 0.127 (0.099) | 0.793 (0.167) | 0.104 (0.129) | 0.125 (0.068) | 0.863 (0.077) | 0.098 (0.074) | 0.079 (0.022) | 0.893 (0.023) | 0.074 (0.021) |
| NOTEARS | 0.792 (0.024) | 0.825 (0.031) | 0.291 (0.027) | 0.761 (0.067) | 0.948 (0.041) | 0.139 (0.052) | 0.538 (0.029) | 0.919 (0.038) | 0.149 (0.030) |
| DAG-GNN | **0.808 (0.051)** | 0.792 (0.043) | 0.322 (0.047) | **0.763 (0.036)** | 0.902 (0.027) | 0.167 (0.030) | 0.573 (0.053) | 0.827 (0.058) | 0.192 (0.055) |
| $n = 500$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.801 (0.117) | **0.232 (0.101)** | **0.759 (0.100)** | **0.818 (0.086)** | **0.196 (0.068)** | **0.793 (0.074)** | **0.854 (0.050)** | **0.160 (0.111)** | **0.843 (0.082)** |
| RFCI | 0.469 (0.051) | 0.529 (0.063) | 0.383 (0.060) | 0.430 (0.028) | 0.548 (0.031) | 0.397 (0.028) | 0.195 (0.020) | 0.834 (0.019) | 0.151 (0.019) |
| RICA | 0.683 (0.109) | 0.748 (0.033) | 0.254 (0.077) | 0.598 (0.060) | 0.884 (0.011) | 0.127 (0.033) | 0.701 (0.032) | 0.945 (0.003) | 0.113 (0.012) |
| CAM | 0.471 (0.179) | 0.906 (0.036) | 0.013 (0.113) | 0.345 (0.059) | 0.926 (0.031) | 0.073 (0.045) | 0.158 (0.018) | 0.968 (0.014) | 0.023 (0.028) |
| GDS | 0.230 (0.030) | 0.791 (0.029) | 0.137 (0.031) | 0.378 (0.025) | 0.673 (0.052) | 0.281 (0.068) | 0.326 (0.063) | 0.671 (0.081) | 0.311 (0.058) |
| RESIT | 0.194 (0.103) | 0.787 (0.091) | 0.129 (0.099) | 0.030 (0.039) | 0.864 (0.051) | 0.074 (0.043) | 0.274 (0.134) | 0.799 (0.075) | 0.213 (0.106) |
| IGCI | 0.167 (0.059) | 0.640 (0.109) | 0.193 (0.076) | 0.191 (0.075) | 0.784 (0.092) | 0.171 (0.085) | 0.096 (0.021) | 0.862 (0.038) | 0.098 (0.029) |
| EMD | 0.111 (0.052) | 0.770 (0.127) | 0.105 (0.083) | 0.200 (0.081) | 0.774 (0.099) | 0.181 (0.093) | 0.079 (0.015) | 0.889 (0.021) | 0.076 (0.017) |
| bQCD | 0.100 (0.035) | 0.795 (0.086) | 0.087 (0.052) | 0.175 (0.080) | 0.807 (0.102) | 0.151 (0.093) | 0.070 (0.016) | 0.900 (0.026) | 0.066 (0.021) |
| NOTEARS | 0.899 (0.013) | 0.827 (0.020) | 0.343 (0.018) | 0.783 (0.033) | 0.918 (0.029) | 0.179 (0.031) | 0.572 (0.024) | 0.913 (0.046) | 0.145 (0.033) |
| DAG-GNN | **0.923 (0.022)** | 0.804 (0.029) | 0.379 (0.025) | 0.815 (0.035) | 0.891 (0.037) | 0.224 (0.035) | 0.590 (0.028) | 0.906 (0.029) | 0.166 (0.029) |
| $n = 1000$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.884 (0.109) | **0.196 (0.082)** | **0.826 (0.098)** | **0.861 (0.065)** | **0.185 (0.043)** | **0.830 (0.052)** | **0.867 (0.044)** | **0.160 (0.082)** | **0.861 (0.065)** |
| RFCI | 0.489 (0.048) | 0.567 (0.039) | 0.365 (0.041) | 0.469 (0.025) | 0.546 (0.025) | 0.418 (0.025) | 0.222 (0.032) | 0.848 (0.023) | 0.152 (0.027) |
| RICA | 0.703 (0.079) | 0.759 (0.023) | 0.243 (0.056) | 0.714 (0.031) | 0.884 (0.005) | 0.147 (0.017) | 0.828 (0.021) | 0.945 (0.002) | 0.129 (0.008) |
| CAM | 0.468 (0.132) | 0.906 (0.026) | 0.012 (0.083) | 0.392 (0.077) | 0.909 (0.038) | 0.096 (0.082) | 0.123 (0.046) | 0.963 (0.014) | 0.027 (0.027) |
| GDS | 0.296 (0.064) | 0.773 (0.039) | 0.175 (0.054) | 0.332 (0.079) | 0.756 (0.071) | 0.195 (0.084) | 0.313 (0.085) | 0.673 (0.077) | 0.319 (0.023) |
| RESIT | 0.216 (0.094) | 0.778 (0.044) | 0.150 (0.069) | 0.032 (0.039) | 0.861 (0.045) | 0.076 (0.042) | 0.375 (0.071) | 0.836 (0.032) | 0.183 (0.058) |
| IGCI | 0.111 (0.117) | 0.750 (0.264) | 0.119 (0.183) | 0.183 (0.059) | 0.813 (0.063) | 0.149 (0.064) | 0.115 (0.021) | 0.836 (0.091) | 0.119 (0.045) |
| EMD | 0.111 (0.117) | 0.750 (0.264) | 0.119 (0.183) | 0.204 (0.029) | 0.793 (0.033) | 0.170 (0.033) | 0.110 (0.009) | 0.865 (0.036) | 0.101 (0.018) |
| bQCD | 0.111 (0.117) | 0.750 (0.764) | 0.119 (0.183) | 0.246 (0.088) | 0.749 (0.095) | 0.214 (0.096) | 0.113 (0.020) | 0.849 (0.080) | 0.114 (0.041) |
| NOTEARS | **0.964 (0.025)** | 0.738 (0.031) | 0.426 (0.030) | 0.836 (0.037) | 0.913 (0.042) | 0.172 (0.035) | 0.678 (0.022) | 0.903 (0.029) | 0.126 (0.024) |
| DAG-GNN | 0.952 (0.037) | 0.722 (0.031) | 0.433 (0.033) | 0.802 (0.039) | 0.822 (0.041) | 0.251 (0.045) | 0.699 (0.028) | 0.857 (0.032) | 0.149 (0.030) |

Table 2: Simulation Scenario 3. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses. The best performance is shown in boldface.

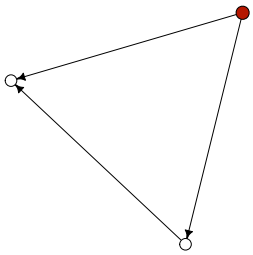
| $n = 125$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.719 (0.063) | **0.281 (0.081)** | **0.657 (0.079)** | 0.712 (0.058) | **0.326 (0.043)** | **0.628 (0.032)** | 0.688 (0.068) | **0.329 (0.039)** | **0.616 (0.048)** |
| LiNG | **0.873 (0.090)** | 0.864 (0.006) | 0.031 (0.038) | **0.875 (0.082)** | 0.917 (0.011) | 0.021 (0.043) | **0.752 (0.094)** | 0.928 (0.014) | 0.009 (0.031) |
| ANM | 0.128 (0.021) | 0.866 (0.049) | 0.029 (0.030) | 0.031 (0.046) | 0.855 (0.042) | 0.016 (0.024) | 0.018 (0.033) | 0.863 (0.048) | 0.011 (0.031) |
| $n = 250$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | 0.813 (0.033) | **0.252 (0.072)** | **0.753 (0.068)** | 0.751 (0.049) | **0.322 (0.055)** | **0.727 (0.058)** | 0.745 (0.056) | **0.322 (0.041)** | **0.725 (0.044)** |
| LiNG | **0.842 (0.073)** | 0.866 (0.009) | 0.025 (0.048) | **0.856 (0.072)** | 0.920 (0.008) | 0.013 (0.042) | **0.768 (0.833)** | 0.953 (0.010) | 0.006 (0.039) |
| ANM | 0.133 (0.043) | 0.851 (0.027) | 0.029 (0.048) | 0.029 (0.020) | 0.917 (0.048) | 0.007 (0.033) | 0.028 (0.032) | 0.855 (0.048) | 0.022 (0.044) |
| $n = 500$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | **0.891 (0.031)** | **0.234 (0.069)** | **0.782 (0.065)** | **0.885 (0.045)** | **0.257 (0.052)** | **0.754 (0.037)** | **0.786 (0.041)** | **0.319 (0.029)** | **0.748 (0.027)** |
| LiNG | 0.809 (0.072) | 0.867 (0.010) | 0.015 (0.049) | 0.823 (0.098) | 0.915 (0.014) | 0.014 (0.030) | 0.743 (0.086) | 0.947 (0.012) | 0.005 (0.038) |
| ANM | 0.138 (0.026) | 0.827 (0.021) | 0.027 (0.036) | 0.021 (0.039) | 0.847 (0.040) | 0.016 (0.041) | 0.022 (0.045) | 0.853 (0.042) | 0.019 (0.038) |
| $n = 1000$ | $p = 10$ | | | $p = 25$ | | | $p = 50$ | | |
| | TPR | FDR | MCC | TPR | FDR | MCC | TPR | FDR | MCC |
| CHOD | **0.953 (0.028)** | **0.219 (0.071)** | **0.844 (0.063)** | **0.947 (0.033)** | **0.247 (0.039)** | **0.839 (0.031)** | **0.939 (0.029)** | **0.251 (0.018)** | **0.835 (0.023)** |
| LiNG | 0.805 (0.073) | 0.855 (0.012) | 0.021 (0.037) | 0.784 (0.075) | 0.916 (0.010) | 0.011 (0.046) | 0.766 (0.087) | 0.933 (0.010) | 0.008 (0.045) |
| ANM | 0.167 (0.028) | 0.856 (0.042) | 0.016 (0.037) | 0.031 (0.023) | 0.849 (0.039) | 0.018 (0.021) | 0.021 (0.033) | 0.877 (0.054) | 0.011 (0.029) |

Table 3: Simulation Scenario 2 with $Z$ included. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses. The best performance is shown in boldface.

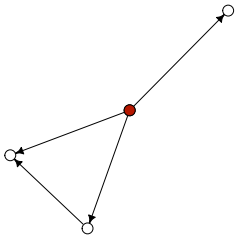
| $n = 125$ | CHOD | RFCI | RICA | CAM | GDS | RESIT | IGCI | EMD | bQCD |
|---|---|---|---|---|---|---|---|---|---|
| TPR | **0.659 (0.114)** | 0.183 (0.065) | 0.003 (0.017) | 0.003 (0.017) | 0.143 (0.074) | 0.040 (0.060) | 0.013 (0.052) | 0.083 (0.042) | 0.080 (0.029) |
| FDR | **0.258 (0.103)** | 0.745 (0.079) | 0.917 (0.289) | 0.993 (0.033) | 0.823 (0.086) | 0.863 (0.232) | 0.960 (0.138) | 0.690 (0.100) | 0.700 (0.093) |
| MCC | **0.656 (0.080)** | 0.112 (0.074) | -0.017 (0.091) | -0.061 (0.020) | 0.057 (0.084) | 0.009 (0.103) | -0.042 (0.084) | 0.107 (0.062) | 0.102 (0.050) |
| $n = 500$ | CHOD | RFCI | RICA | CAM | GDS | RESIT | IGCI | EMD | bQCD |
| TPR | **0.801 (0.117)** | 0.266 (0.057) | 0.000 (0.000) | 0.208 (0.102) | 0.195 (0.072) | 0.257 (0.096) | 0.190 (0.051) | 0.202 (0.071) | 0.178 (0.061) |
| FDR | **0.232 (0.101)** | 0.738 (0.051) | 1.000 (0.000) | 0.703 (0.139) | 0.789 (0.065) | 0.833 (0.061) | 0.719 (0.078) | 0.709 (0.063) | 0.822 (0.078) |
| MCC | **0.759 (0.100)** | 0.142 (0.058) | -0.053 (0.020) | 0.166 (0.128) | 0.100 (0.070) | 0.068 (0.086) | 0.147 (0.067) | 0.159 (0.071) | 0.081 (0.073) |



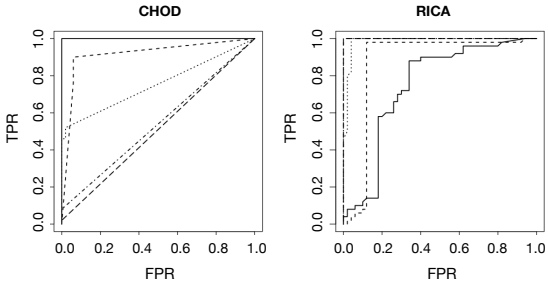

(a) Direct causal effect functions.



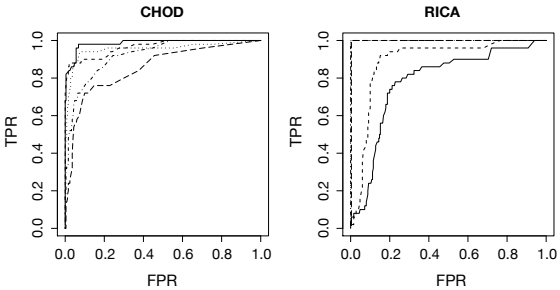
(b) The three-node graph.



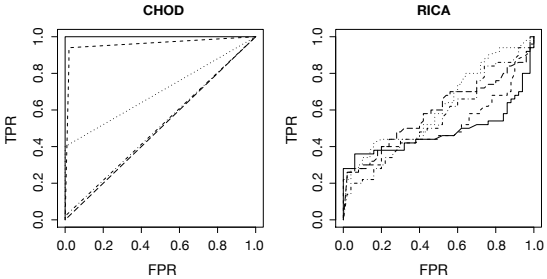
(c) The four-node graph.



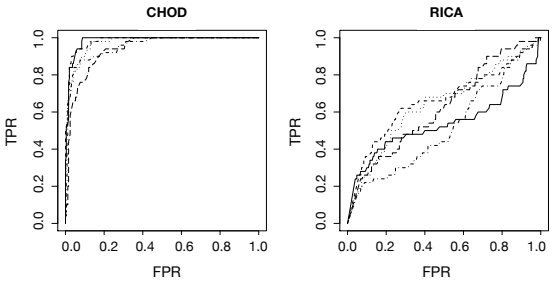

(d) Three-node graph with uniform noises.




(e) Four-node graph with uniform noises.

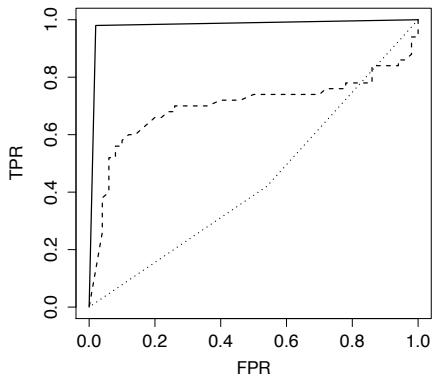



(f) Three-node graph with Gaussian noises.

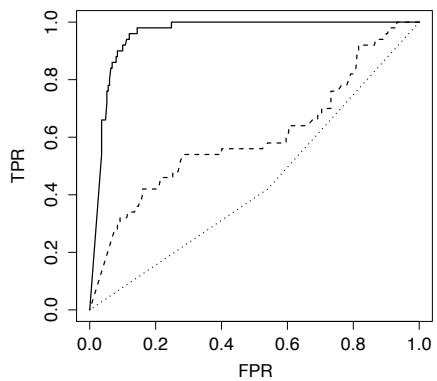



(g) Four-node graph with Gaussian noises.

Figure 5: Misspecification 1. (a) Simulation true direct causal effect functions. (b)&(c) Simulation true graphs. Solid red nodes are latent (discarded at model fitting stage). (d)&(g) Receiver operating characteristics curves for recovering causal relationships between observed variables under varying degrees of heterogeneity are represented by the same line types as shown in (a).

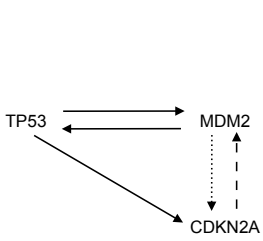



(a) Three-node graph.

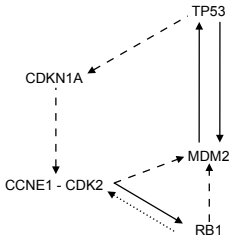



(b) Four-node graph.

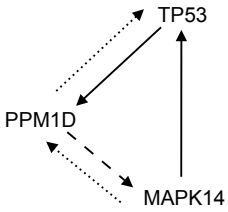
Figure 6: Misspecified model 2. Receiver operating characteristics curves for CHOD, RICA, and CAM are represented by solid, dashed, and dotted lines, respectively.



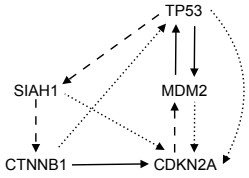

(a) Network A.



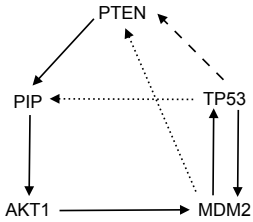

(b) Network B.



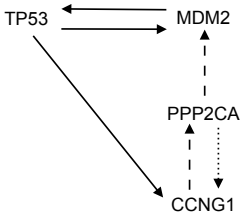

(c) Network C.



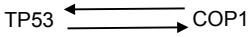

(d) Network D.



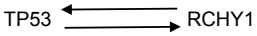

(e) Network E.




(f) Network F.




(g) Network G.




(h) Network H.

Figure 7: Estimated feedback loops using the proposed CHOD. Solid arrows are true positives, dashed arrows are false negatives, and dotted arrows are false positives.