Causal Discovery with Heterogeneous Observational Data: Supplementary Materials

Fangting Zhou[1,2] **Kejun He**[2] **Yang Ni**[1]

[1]Department of Statistics, Texas A&M University, College Station, Texas, USA
[2]Institute of Statistics and Big Data, Renmin University of China, Beijing, China

S1 COMPARISONS WITH HETEROGENEOUS CAUSAL DISCOVERY METHODS

We notice that our method shares some similarities with JCI [Mooij et al., 2020] and CD-NOD [Huang et al., 2020] in establishing causal identifiability with the help of heterogeneous data/environments, indicated by the variables Z. However, how Z enters the causal model is very different, which in turn leads to significant methodological, theoretical, and computational differences.

Causal identification While JCI and CD-NOD are flexible in utilizing environment information and dealing with various kinds of distributions by learning the graph jointly over Z and X through conditional independence tests, there are many situations where the causal structure is only partially identifiable. For example, consider two competing causal models

$$M_1 : X_1 = \epsilon_1,\ \ X_2 = B_{21}(Z)X_1 + \epsilon_2,\ \ X_3 = B_{31}(Z)X_1 + B_{32}(Z)X_2 + \epsilon_3,$$
$$M_2 : X_1 = \epsilon_1,\ \ X_2 = B^{\star}_{21}(Z)X_1 + B^{\star}_{23}(Z)X_3 + \epsilon_2,\ \ X_3 = B^{\star}_{31}(Z)X_1 + \epsilon_3.$$

Their corresponding causal graphs of X_1, X_2, X_3, and Z are the same except that the arrow direction between X_2 and X_3 is reversed. Because the graphs are Markov equivalent, the causal direction between X_2 and X_3 cannot be identified by JCI and CD-NOD. On the contrary, our method is able to identify the direction as established by our theorems.

Proof techniques Because of the difference illustrated in the example above, to prove our causal identifiability results, existing proofs in the literature do not apply and significant efforts are needed to figure out how heterogeneity helps structure learning with our model formulation (through varying causal effects). JCI and CD-NOD are able to narrow down the Markov equivalence class under general assumptions like faithfulness, but there could still be causal indeterminacy without additional assumptions. For example, if one would like to assume the observations are subject to a diverse set of hard interventions, then one still has to make assumptions on the interventional experiments to fully identify causal structures [Hyttinen et al., 2013].

Environment variable To our knowledge, JCI mainly focuses on observations from a finite number of contexts (i.e., Z is discrete). On the contrary, we allow Z to be continuous, i.e., the environment can change continuously and may be different for each observation. In other words, we can have n environments, one for each of the n observations. For JCI to be applicable to continuous Z, one possible solution is to discretize Z, but the causal identification may be sensitive to the method of discretization. Our parameterization naturally allows borrowing of information from observations in similar environments thus overcomes this problem. While CD-NOD allows continuous environments, it is not clear how to generalize it to allow cycles and confounders.

Algorithm JCI and CD-NOD are constraint-based methods relying on conditional independence tests, which are known to lack statistical power even just for a moderately large conditioning set (say, 10) and require a large sample size and careful adjustment for multiplicity. By contrast, our method is fully model-based and hence is significantly less prone to the curse of dimensionality. In addition, to the best of our knowledge, current ASD-JCI version allows no more than $p = 10$ variables, largely limited by the statistical power and the computational burden of constraint-based causal discovery.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Additional simulations We now compared the performance of ASD-JCI and the method from Faria et al. [2022] with our proposed method. We considered two scenarios. In the first scenario, we generated $n = 200$ samples from the following model

$$\begin{aligned}
&X_1 = \epsilon_1,\ X_2 = \epsilon_2,\\
&X_3 = 0.8Z \times X_1 + 0.5Z \times X_2 + \epsilon_3,\\
&X_4 = -0.9Z \times X_1 - 0.5Z \times X_2 + 0.9Z \times X_3 + \epsilon_4,
\end{aligned}$$

where $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \sim N(0,1)$ and $Z \sim U(-1,1)$. We assumed X_2 was not observed at the model fitting stage and served as a latent confounder between X_3 and X_4. The underlying causal model on $X = (X_1, X_3, X_4)$ was acyclic and causally insufficient. Notice that in this case, the confounding was not stable and independent of Z. We applied ASD-JCI123 (one version of JCI which had the top performance for unknown interventional targets in the experiments of their paper) with acyclic = TRUE, sufficient = FALSE, test = gaussCItest (i.e., the d-separation criterion from Hyttinen et al. [2014]). Other parameters were set to their default values. The comparison with our method is shown in Table S1 where our method outperformed the other two competitors. The method of Faria et al. [2022] performed worst in this example, since it applies to discrete environments and it remains unclear how the clustering is carried out in the continuous case.

Table S1: Additional simulation experiment – acyclic model. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses.

CHOD			ASD-JCI			Faria et al.		
TPR	FDR	MCC	TPR	FDR	MCC	TPR	FDR	MCC
0.860 (0.203)	0.120 (0.199)	0.805 (0.296)	0.753 (0.284)	0.380 (0.126)	0.518 (0.273)	0.707 (0.145)	0.480 (0.069)	0.363 (0.139)

In the second scenario, we generated data from a cyclic model

$$\begin{aligned}
&X_1 = 0.8Z \times X_3 + \epsilon_1,\ X_2 = 0.9\cos(\pi Z) \times X_1 + \epsilon_2,\\
&X_3 = 0.9\tanh(\pi Z) \times X_2 + \epsilon_3,\ X_4 = -0.8Z \times X_3 + \epsilon_4,\ X_5 = 0.9\sin(\pi Z) \times X_4 + \epsilon_5,
\end{aligned}$$

where X_1, X_2, X_3 form a cycle. We compared with ASD-JCI123 and set acyclic = FALSE, sufficient = TRUE, test = gaussCItest (i.e., the σ-separation criterion from Forré and Mooij [2018]). Other parameters were set to their default values. The results are shown in Table S2 where CHOD still significantly outperformed ASD-JCI123, which had a high FDR.

Table S2: Additional simulation experiment – cyclic model. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses.

CHOD			ASD-JCI123		
TPR	FDR	MCC	TPR	FDR	MCC
0.860 (0.172)	0.206 (0.102)	0.758 (0.161)	0.820 (0.063)	0.758 (0.019)	0.150 (0.069)

S2 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 non-linear 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 $pa^+(j) = \{j\} \cup pa(j)$.

S3 PROOFS

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

S3.1 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,

$$\begin{aligned}
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{11} b_{21}^2(Z) + 2\sigma_{12} b_{21}(Z) + \sigma_{22}) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{11} b'^2_{21}(Z) + 2\sigma'_{12} b'_{21}(Z) + \sigma'_{22}),\\
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{22} b_{12}^2(Z) + 2\sigma_{12} b_{12}(Z) + \sigma_{11}) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{22} b'^2_{12}(Z) + 2\sigma'_{12} b'_{12}(Z) + \sigma'_{11}),\\
&(\sigma'_{11}\sigma'_{22} - \sigma'^2_{12})(\sigma_{11} b_{21}(Z) + \sigma_{22} b_{12}(Z) + \sigma_{12} b_{12}(Z) b_{21}(Z) + \sigma_{12})\\
&\qquad = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma'_{11} b'_{21}(Z) + \sigma'_{22} b'_{12}(Z) + \sigma'_{12} b'_{12}(Z) b'_{21}(Z) + \sigma'_{12}).
\end{aligned} \tag{S1}$$

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 (S1) 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, $\mathcal{G} = \mathcal{G}'$. □

S3.2 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

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

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 jth node, when $\{ds_{[j]}(j) \cup pa(ds_{[j]}(j) \cup \{j\})\} \subseteq C \subseteq [j-1]$,

$$\mathrm{Cov}(X_k, X_j | \boldsymbol{X}_S) = \mathrm{Cov}(\textstyle\sum_{\ell \notin ds_{[j]}(j)} b_{\ell \to k}(Z)\epsilon_\ell + \epsilon_k, \epsilon_j | \boldsymbol{X}_C) = 0, \ \forall j > k \notin C, \tag{S2}$$

where $b_{\ell \to k}(Z) = [(\boldsymbol{I} - \boldsymbol{B}(Z))^{-1}]_{k\ell}$ is the total causal effect from X_ℓ 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 jth node [Richardson, 2003]. We take the minimum set for which the conditional independence condition (S2) 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 jth 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_ℓ to X_j, i.e., X_ℓ 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 $[p]\setminus S$ 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.

S3.3 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 $c^\ell_r \to c^j_t$ for some $c^\ell_r \in C_\ell$ and $c^j_t \in C_j$. 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 $c^\ell_1 \to \cdots \to c^\ell_{|C_\ell|} \to c^\ell_1$ forms the maximal cyclic component C_ℓ. 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, $\boldsymbol{A}_{D,D}(Z) = (\boldsymbol{I} - \boldsymbol{B}_{D,D}(Z))^{-1}$, $\boldsymbol{E}_{D,C} = (\boldsymbol{e}^T_1, \ldots, \boldsymbol{e}^T_{|D|})^T$ 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

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

such that $\mathrm{Cov}(\boldsymbol{A}^{'-1}_{D,D}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}) = \mathrm{diag}(\sigma'_{d_1}, \ldots, \sigma'_{d_{|D|}})$. 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 $\boldsymbol{A}^{'-1}_{D,D}(Z)$ such that $\mathrm{Cov}(\boldsymbol{A}^{'-1}_{D,D}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}})$ is a diagonal matrix.

Notice that when $D = C_\ell$ which is a validate candidate, we can choose $\boldsymbol{A}^{'-1}_{D,D}(Z) = \boldsymbol{A}^{-1}_{C_\ell,C_\ell}(Z)$ which leads to $\mathrm{Cov}(\boldsymbol{A}^{'-1}_{D,D}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}) = \mathrm{Cov}(\boldsymbol{\epsilon}_{C_\ell}) = \mathrm{diag}(\sigma_{c^\ell_1}, \ldots, \sigma_{c^\ell_{|C_\ell|}})$; hence C_ℓ 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,

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

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 $\mathrm{Cov}(\boldsymbol{A}^{'-1}_{D,D}(Z)\boldsymbol{X}_D|\boldsymbol{X}_{C_1}, \ldots, \boldsymbol{X}_{C_{\ell-1}}) = \mathrm{diag}(\sigma'_{d_1}, \ldots, \sigma'_{d_{|D|}})$ 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,

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

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

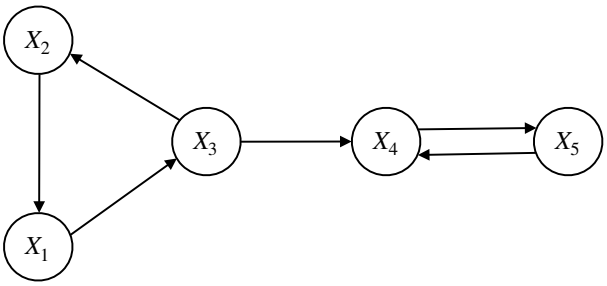


Figure S1: A demonstrative example of Theorem 3.

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 S1. The maximal cyclic components are $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5\}$. The collapsed graph $\widetilde{\mathcal{G}}$ is simply $C_1 \to C_2$.

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 $\boldsymbol{A}'^{-1}_{D,D}$ such $\mathrm{Cov}(\boldsymbol{A}'^{-1}_{D,D}(Z)\boldsymbol{X}_D)$ 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

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

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.

S3.4 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. □

S4 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 τ from the full conditional distribution

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

5. We generate the edge inclusion probability π from the full conditional distribution

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

S4.1 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

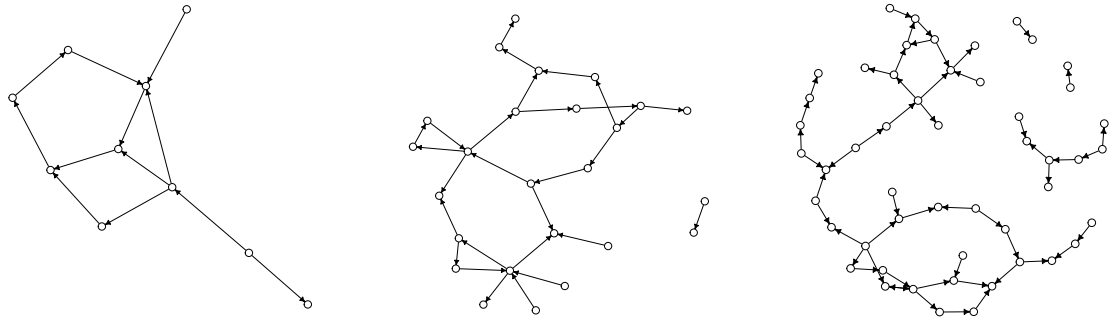
$$\mathbb{P}(z_1, \ldots, z_n) \propto \textstyle\prod_{j=i+1}^{n} \sin^{2\gamma}\{\pi(z_i - z_j)\},\ \forall z_i \in [0, 1]$$

with the repulsive parameter γ. 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

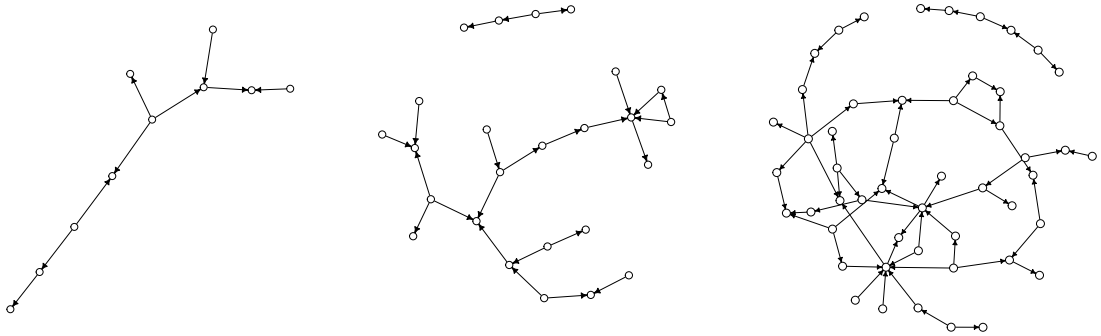
$$\begin{aligned}\log \alpha = {} & \log\{\mathbb{Q}(z_i|z'_i)\mathbb{P}(z'_i, \boldsymbol{z}_{-i})\mathbb{P}(\boldsymbol{x}_i|z'_i, \boldsymbol{B}(z'_i), \boldsymbol{S})\} \\ & - \log\{\mathbb{Q}(z'_i|z_i)\mathbb{P}(z_i, \boldsymbol{z}_{-i})\mathbb{P}(\boldsymbol{x}_i|z_i, \boldsymbol{B}(z_i), \boldsymbol{S})\}.\end{aligned}$$

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



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

Figure S2: Simulation true graphs in Scenario 1 (cyclic graphs with confounders).



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

Figure S3: Simulation true graphs in Scenario 2 (acyclic graphs with confounders).

S5 ADDITIONAL DETAILS OF THE EXPERIMENTS

Figures S2–S4 show the randomly generated simulation true causal graphs in Scenarios 1–3.

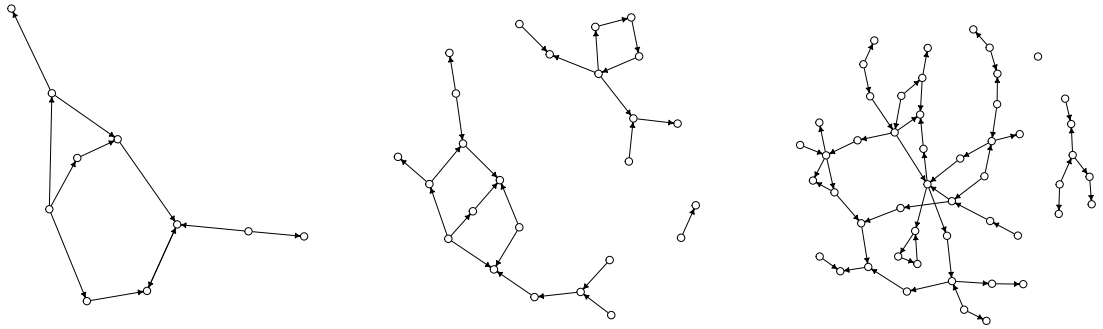
S5.1 ADDITIONAL DETAILS OF SIMULATION SCENARIO 2 AND 3

Table S3 and S4 respectively show summaries for simulation scenario 2 and 3. Clearly, CHOD outperformed others by a significant amount. Further, we compared our method with CAM, RESIT, IGCI, EMD, bQCD, NOTEARS, and DAG-GNN in the acyclic graphs without confounders scenario. The result is shown in Table S5. However, the performance of these methods did not improve much compared to the scenario with confounders, and the proposed CHOD still significantly outperformed them. We suspect this is because the simulated data are heterogeneous and these methods were not designed to handle data heterogeneity. Additionally, we used $p = 10$ and $n \in \{125, 500\}$ in the acyclic graph with confounders case 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 S6. The conclusion stays the same: CHOD outperforms the alternatives with larger TPR and smaller FDR.

S5.2 ADDITIONAL DETAILS OF MODEL MISSPECIFICATION

Misspecification 1 Figure S5 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



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

Figure S4: Simulation true graphs in Scenario 3 (cyclic graphs without confounders).

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 S6. 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.

S5.3 ADDITIONAL RESULTS FOR THE APPLICATION

The estimated networks from CHOD are shown in Figure S7. CHOD performed especially well on the PTEN/AKT/MDM-2 loop (Network E).

References

Gonçalo RA Faria, André FT Martins, and Mário AT Figueiredo. Differentiable causal discovery under latent interventions. *arXiv preprint arXiv:2203.02336*, 2022.

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.

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.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

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.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 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 S3: 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 S4: 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.

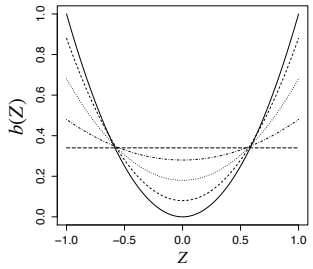
$n = 125$	$p = 10$			$p = 25$			$p = 50$		
	TPR	FDR	MCC	TPR	FDR	MCC	TPR	FDR	MCC
CHOD	0.719 (0.063)	**0.281 (0.081)**	**0.657 (0.079)**	0.712 (0.058)	**0.326 (0.043)**	**0.628 (0.032)**	0.688 (0.068)	**0.329 (0.039)**	**0.616 (0.048)**
LiNG	**0.873 (0.090)**	0.864 (0.006)	0.031 (0.038)	**0.875 (0.082)**	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 S5: Simulation acyclic graph without confounders. $n = 500$ and $p = 10$. Average operating characteristics over 50 repetitions. The standard deviation for each statistic is given within parentheses. The best performance is shown in boldface.

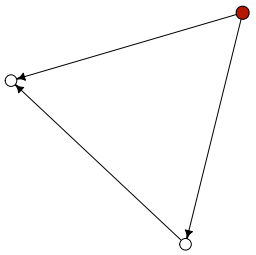
	CHOD	CAM	RESIT	IGCI	EMD	bQCD	NOTEARS	DAG-GNN
TPR	0.759 (0.141)	0.068 (0.061)	0.078 (0.054)	0.178 (0.099)	0.055 (0.008)	0.112 (0.009)	0.843 (0.021)	**0.855 (0.019)**
FDR	**0.224 (0.142)**	0.811 (0.207)	0.697 (0.308)	0.468 (0.298)	0.921 (0.003)	0.678 (0.011)	0.692 (0.033)	0.652 (0.028)
MCC	**0.743 (0.152)**	0.066 (0.105)	0.107 (0.121)	0.272 (0.183)	0.002 (0.001)	0.149 (0.009)	0.475 (0.027)	0.481 (0.029)

Table S6: Simulation 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.

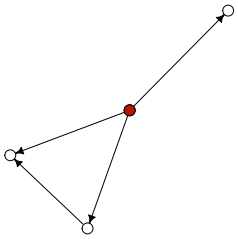
$n = 125$	CHOD	RFCI	RICA	CAM	GDS	RESIT	IGCI	EMD	bQCD
TPR	**0.659 (0.114)**	0.183 (0.065)	0.003 (0.017)	0.003 (0.017)	0.143 (0.074)	0.040 (0.060)	0.013 (0.052)	0.083 (0.042)	0.080 (0.029)
FDR	**0.258 (0.103)**	0.745 (0.079)	0.917 (0.289)	0.993 (0.033)	0.823 (0.086)	0.863 (0.232)	0.960 (0.138)	0.690 (0.100)	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)



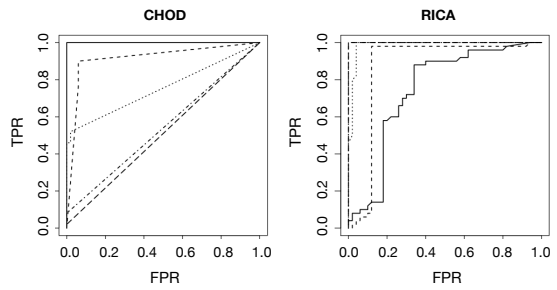
(a) Direct causal effect functions.



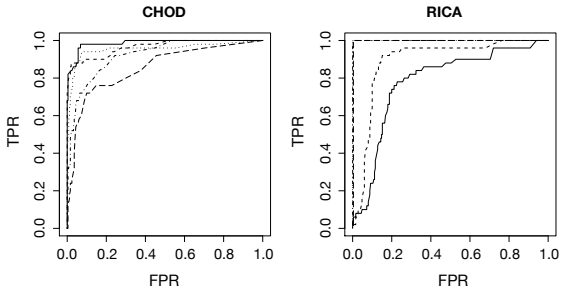
(b) The three-node graph.



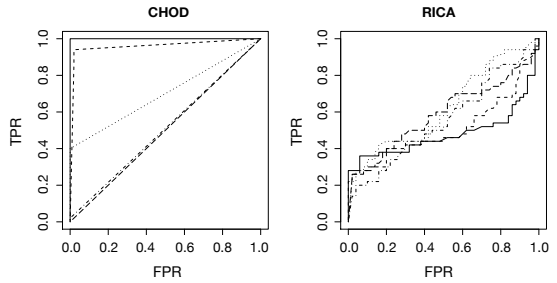
(c) The four-node graph.



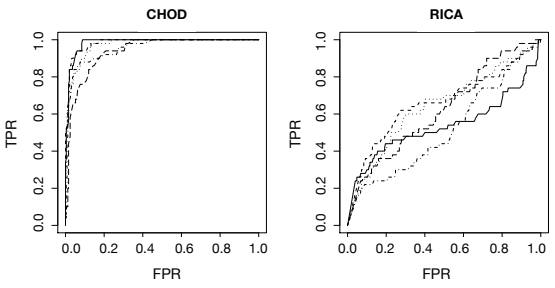
(d) Three-node graph with uniform noises.



(e) Four-node graph with uniform noises.

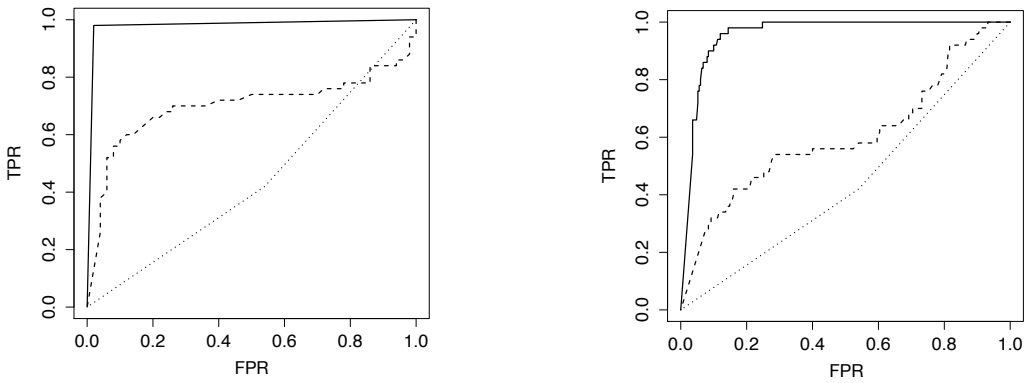


(f) Three-node graph with Gaussian noises.



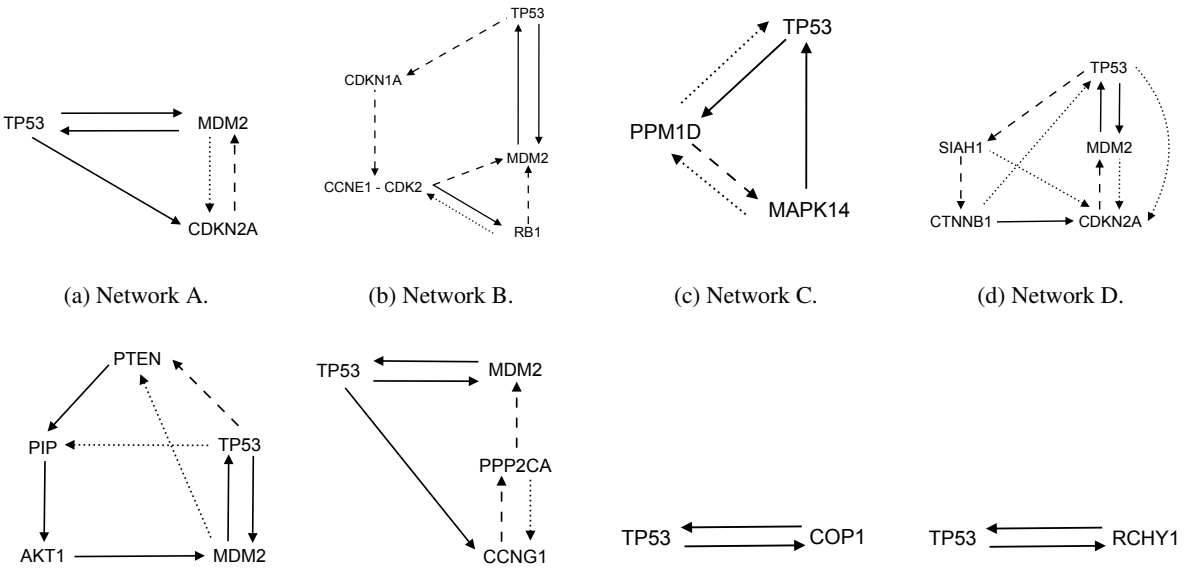
(g) Four-node graph with Gaussian noises.

Figure S5: 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 S6: Misspecification 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 S7: Estimated feedback loops using the proposed CHOD. Solid arrows are true positives, dashed arrows are false negatives, and dotted arrows are false positives.