Appendix for

"Federated Causal Discovery from Heterogeneous Data"

Appendix organization:

A1 RELATED WORKS

Causal Discovery. In general, there are mainly three categories of methods for causal discovery (CD) from observed data (Spirtes & Zhang, 2016): constraint-based methods, score-based methods and function-based methods. Constraint-based methods utilize the conditional independence test (CIT) to learn a skeleton of the directed acyclic graph (DAG), and then orient the edges upon the skeleton. Such methods contain Peter-Clark (PC) algorithm (Spirtes & Zhang, 2016) and Fast Causal Inference (FCI) algorithm (Spirtes, 2001). Some typical CIT methods include kernel-based independent conditional test (Zhang et al., 2012) and approximate kernel-based conditional independent test (Strobl et al., 2019). Score-based methods use a score function and a greedy search method to learn a DAG with the highest score by searching all possible DAGs from the data, such as Greedy Equivalent Search (GES) (Chickering, 2002). Within the score-based category, there is a continuous optimization-base subcategory attracting increasing attention. NOTEARS (Zheng et al., 2018) firstly reformulates the DAG learning process as a continuous optimization problem and solves it using gradient-based method. NOTEARS is designed under the assumption of the linear relations between variables. Subsequent works have extended NOTEARS to handle nonlinear cases via deep neural networks, such as DAG-GNN (Yu et al., 2019) and DAG-NoCurl (Yu et al., 2021). ENCO (Lippe et al., 2022) presents an efficient DAG discovery method for directed acyclic causal graphs utilizing both observational and interventional data. AVCI (Lorch et al., 2022) infers causal structure by performing amortized variational inference over an arbitrary data-generating distribution. These continuous-optimization-based methods might suffer from various technical issues, including convergence (Wei et al., 2020; Ng et al., 2022), nonconvexity (Ng et al., 2023), and sensitivity to data standardization (Reisach et al., 2021). Function-based methods rely on the causal asymmetry property, including the linear non-Gaussion model (LiNGAM) (Shimizu et al., 2006), the additive noise model (Hoyer et al., 2008), and the post-nonlinear causal model (Zhang & Hyvarinen, 2012).

Causal Discovery from Heterogeneous Data. Most of the causal discovery methods mentioned above usually assume that the data is independently and identically distributed (i.i.d.). However, in practical scenarios, distribution shift is possibly occurring across datasets, which can be changing across different domains or over time, as featured by heterogeneous or non-stationary data (Huang et al., 2020). To tackle the issue of changing causal models, one may try to find causal models on sliding windows for non-stationary data (Calhoun et al., 2014), and then compare them. Improved versions include the regime aware learning algorithm to learn a sequence of Bayesian networks that model a system with regime changes (Bendtsen, 2016). Such methods may suffer from high estimation variance due to sample scarcity, large type II errors, and a large number of statistical tests. Some methods aim to estimate the time-varying causal model by making use of certain types of smoothness of the change (Huang et al., 2015), but they do not explicitly locate the changing causal modules. Several methods aim to model time-varying time-delayed causal relations (Xing et al., 2010), which can be reduced to online parameter learning because the direction of the causal relations is given (i.e., the past influences the future). Moreover, most of these methods assume linear causal models, limiting their applicability to complex problems with nonlinear causal relations. In particular, a nonparametric constraint-based method to tackle this causal discovery problem from non-stationary or heterogeneous data, called CD-NOD (Huang et al., 2020), was recently proposed, where the surrogate variable was introduced, written as smooth functions of time or domain index. The first model-based method was proposed for heterogeneous data in the presence of cyclic causality and confounders, named CHOD (Zhou et al., 2022). Saeed et al. (Saeed et al., 2020) provided a graphical representation via the mixture DAG of distributions that arise as mixtures of causal DAGs.

Federated Causal Discovery. A two-step procedure was adopted (Gou et al., 2007) to learn a DAG from horizontally partitioned data, which firstly estimated the structures independently using each client's local dataset, and secondly applied further conditional independence test. Instead of using statistical test in the second step, a voting scheme was used to pick those edges identified by more than half of the clients (Na & Yang, 2010). These methods leverage only the final graphs independently estimated from each local dataset, which may lead to suboptimal performance as the information exchange may be rather limited. Furthermore, (Samet & Miri, 2009) developed a privacy-preserving method based on secure multiparty computation, but was limited to the discrete case. For vertically partitioned data, (Yang et al., 2019) constructed an approximation to the score function in the discrete case and adopted secure multiparty computation. (Chen et al., 2003) developed a four-step procedure that involves transmitting a subset of samples from each client to a central site, which may lead to privacy concern. NOTEARS-ADMM (Ng & Zhang, 2022) and Fed-DAG

(Gao et al., 2022) were proposed for the federated causal discovery (FCD) based on continuous optimization methods. Fed-PC (Huang et al., 2022) was developed as a federated version of classical PC algorithm, however, it was developed for homogeneous data, which may lead to poor performance on heterogeneous data. DARLIS (Ye et al., 2022) utilizes the distributed annealing (Arshad & Silaghi, 2004) strategy to search for the optimal graph, while PERI (Mian et al., 2023) aggregates the results of the local greedy equivalent search (GES) (Chickering, 2002) and chooses the worst-case regret for each iteration. Fed-CD (Abyaneh et al., 2022) was proposed for both observational and interventional data based on continuous optimization. FEDC^2SL (Wang et al., 2023) extended χ^2 test to the federated version, however, this method is restrictive on discrete variables and therefore not applicable for any continuous variables. Notice that most of these above-mentioned methods heavily rely on either identifiable functional causal models or homogeneous data distributions. These assumptions may be overly restrictive and difficult to be satisfied in real-world scenarios, limiting their diverse applicability.

A2 Details about the Characterization

A2.1 Characterization of Conditional Independence

In this section, we will provide more details about the interpretation of $\Sigma_{\ddot{X}Y|Z}$ as formulated in Eq. 13, the definition of characteristic kernel as shown in Lemma 9, which is helpful to understand the Lemma 1 in the main paper. We then provide the uncorrelatedness-based characterization of CI in Lemma 10.

First of all, for the random vector (X, Y) on $\mathcal{X} \times \mathcal{Y}$, the cross-covariance operator from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}_\mathcal{X}$ is defined by the relation

$$\langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)], \tag{11}$$

for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. Furthermore, we define the partial cross-covariance operator as

$$\Sigma_{XY|Z} = \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}. \tag{12}$$

If Σ_{ZZ} is not invertible, use the right inverse instead of the inverse. We can intuitively interpret the operator $\Sigma_{XY|Z}$ as the partial cross-covariance between $\{f(X), \forall f{\in}\mathcal{H}_\mathcal{X}\}$ and $\{g(Y), \forall g{\in}\mathcal{H}_\mathcal{Y}\}$ given $\{q(Z), \forall q{\in}\mathcal{H}_\mathcal{Z}\}$.

Lemma 9 (Characteristic Kernel (Fukumizu et al., 2007))**.** *A kernel $\mathcal{K}_\mathcal{X}$ is characteristic, if the condition $\mathbb{E}_{X\sim\mathbb{P}_X}[f(X)]{=}\mathbb{E}_{X\sim\mathbb{Q}_X}[f(X)]$ ($\forall f{\in}\mathcal{H}_\mathcal{X}$) implies $\mathbb{P}_X{=}\mathbb{Q}_X$, where $\mathbb{P}_X$ and $\mathbb{Q}_X$ are two probability distributions of X. Gaussian kernel and Laplacian kernel are characteristic kernels.*

As shown in Lemma 1, if we use characteristic kernel and define $\ddot{X} \triangleq (X, Z)$, the characterization of CI could be related to the partial cross-covariance as $\Sigma_{\ddot{X}Y|Z} = 0 \iff X \perp\!\!\!\perp Y|Z$, where

$$\Sigma_{\ddot{X}Y|Z} = \Sigma_{\ddot{X}Y} - \Sigma_{\ddot{X}Z}\Sigma_{ZZ}^{-1}\Sigma_{ZY}. \tag{13}$$

Similarly, we can intuitively interpret the operator $\Sigma_{\ddot{X}Y|Z}$ as the partial cross-covariance between $\{f(\ddot{X}), \forall f{\in}\mathcal{H}_{\ddot{\mathcal{X}}}\}$ and $\{g(Y), \forall g{\in}\mathcal{H}_\mathcal{Y}\}$ given $\{q(Z), \forall q{\in}\mathcal{H}_\mathcal{Z}\}$.

Based on Lemma 1, we further consider a different characterization of CI which enforces the uncorrelatedness of functions in suitable spaces, which may be intuitively more appealing. Denote the probability distribution of X as $\mathbb{P}_X$ and the joint distribution of (X, Y) as $\mathbb{P}_{XY}$. Let L^2_X be the space of square integrable functions of X and L^2_{XY} be that of (X, Y). Specifically, $L^2_X = \{f(X)|\mathbb{E}(f^2) < \infty\}$, and likewise for L^2_{XY}. Particularly, consider the following constrained L^2 spaces:

$$\begin{aligned}
\mathcal{S}_{\ddot{X}} &\triangleq \{f \in L^2_{\ddot{X}} \mid \mathbb{E}(f|Z) = 0\},\\
\mathcal{S}_{\ddot{Y}} &\triangleq \{g \in L^2_{\ddot{Y}} \mid \mathbb{E}(g|Z) = 0\},\\
\mathcal{S}'_{Y|Z} &\triangleq \{g' \mid g' = g(Y) - \mathbb{E}(g|Z), g \in L^2_Y\}.
\end{aligned} \tag{14}$$

They can be constructed from the corresponding L^2 spaces via nonlinear regression. From example, for any function $f \in L^2_{XZ}$, the corresponding function f' is given by:

$$f'(\ddot{X}) = f(\ddot{X}) - \mathbb{E}(f|Z) = f(\ddot{X}) - \beta^*_f(Z), \tag{15}$$

where $\beta^*_f(Z) \in L^2_Z$ is the regression function of $f(\ddot{X})$ on Z. Then, we can then relate the different characterization of CI from Lemma 1 to the uncorrelatedness in the following lemma.

Lemma 10 (Characterization of CI based on Partial Association (Daudin, 1980)). *Each of the following conditions are equivalent to $X \perp\!\!\!\perp Y|Z$*

$$\begin{aligned}
(i.)\ \mathbb{E}(fg) = 0, \forall f \in \mathcal{S}_{\ddot{X}} \text{ and } \forall g \in \mathcal{S}_{\ddot{Y}}, \\
(ii.)\ \mathbb{E}(fg') = 0, \forall f \in \mathcal{S}_{\ddot{X}} \text{ and } \forall g' \in \mathcal{S}'_{Y|Z}, \\
(iii.)\ \mathbb{E}(f\tilde{g}) = 0, \forall f \in \mathcal{S}_{\ddot{X}} \text{ and } \forall \tilde{g} \in L^2_{\ddot{Y}}, \\
(iv.)\ \mathbb{E}(f\tilde{g}') = 0, \forall f \in \mathcal{S}_{\ddot{X}} \text{ and } \forall \tilde{g}' \in L^2_{Y}.
\end{aligned} \tag{16}$$

When (X, Y, Z) are jointly Gaussian, the independence is equivalent to the uncorrelatedness, in other words, $X \perp\!\!\!\perp Y|Z$ is equivalent to the vanishing of the partial correlation coefficient $\rho_{XY|Z}$. We can regard the Lemma 10 as as a generalization of the partial correlation based characterization of CI. For example, condition (i) means that any "residual" function of (X, Z) given Z is uncorrelated with that of (Y, Z) given Z. Here we can observe the similarity between Lemma 1 and Lemma 10, except the only difference that Lemma 10 considers all functions in L^2 spaces, while Lemma 1 exploits the spaces corresponding to some characteristic kernels. If we restrict the function f and g' in condition (ii) to the spaces $\mathcal{H}_{\ddot{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively, Lemma 10 is then reduced to Lemma 1.

Based on the two lemmas mentioned above plus the Lemma 1, we could further derive Lemma 3 in our main paper.

A2.2 Characterization of Independent Change

In Lemma 2 of the main paper, we provide the independent change principle (ICP) to evaluate the dependence between two changing causal models. Here, we give more details about the definition and the assigned value of normalized HSIC. A smaller value means being more independent.

Definition 1 (Normalized HSIC (Fukumizu et al., 2007)). *Given variables U and V, HSIC provides a measure for testing their statistical independence. An estimator of normalized HSIC is given as*

$$\text{HSIC}^{\mathcal{N}}_{UV} = \frac{\text{tr}(\tilde{\boldsymbol{M}}_U \tilde{\boldsymbol{M}}_V)}{\text{tr}(\tilde{\boldsymbol{M}}_U)\,\text{tr}(\tilde{\boldsymbol{M}}_V)}, \tag{17}$$

where $\tilde{\boldsymbol{M}}_U$ and $\tilde{\boldsymbol{M}}_V$ are the centralized Gram matrices, $\tilde{\boldsymbol{M}}_U \triangleq \boldsymbol{H}\boldsymbol{M}_U\boldsymbol{H}$, $\tilde{\boldsymbol{M}}_V \triangleq \boldsymbol{H}\boldsymbol{M}_V\boldsymbol{H}$, $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, $\boldsymbol{I}$ is $n \times n$ identity matrix and $\mathbf{1}$ is vector of n ones. How to construct $\boldsymbol{M}_U$ and $\boldsymbol{M}_V$ will be explained in the corresponding cases below. To check whether two causal modules change independently across different domains, the dependence between $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$ and the dependence between $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$ on the given data can be given by

$$\triangle_{X\to Y} = \frac{\text{tr}(\tilde{\boldsymbol{M}}_X \tilde{\boldsymbol{M}}_{Y|X})}{\text{tr}(\tilde{\boldsymbol{M}}_X)\,\text{tr}(\tilde{\boldsymbol{M}}_{Y|X})}, \quad \triangle_{Y\to X} = \frac{\text{tr}(\tilde{\boldsymbol{M}}_Y \tilde{\boldsymbol{M}}_{X|Y})}{\text{tr}(\tilde{\boldsymbol{M}}_Y)\,\text{tr}(\tilde{\boldsymbol{M}}_{X|Y})}. \tag{18}$$

According to CD-NOD (Huang et al., 2020), instead of working with conditional distribution $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y|X)$, we could use the "joint distribution" $\mathbb{P}(X, Y)$, which is simpler, for estimation. Here we use $\underline{Y}$ instead of Y to emphasize that in this constructed distribution X and Y are not symmetric. Then, the dependence values listed in Eq. 18 could be estimated by

$$\hat{\triangle}_{X\to Y} = \frac{\text{tr}(\tilde{\boldsymbol{M}}_X \tilde{\boldsymbol{M}}_{\underline{Y}X})}{\text{tr}(\tilde{\boldsymbol{M}}_X)\,\text{tr}(\tilde{\boldsymbol{M}}_{\underline{Y}X})}, \quad \hat{\triangle}_{Y\to X} = \frac{\text{tr}(\tilde{\boldsymbol{M}}_Y \tilde{\boldsymbol{M}}_{\underline{X}Y})}{\text{tr}(\tilde{\boldsymbol{M}}_Y)\,\text{tr}(\tilde{\boldsymbol{M}}_{\underline{X}Y})}, \tag{19}$$

where $\tilde{\boldsymbol{M}}_X \triangleq \boldsymbol{H}\boldsymbol{M}_X\boldsymbol{H}$, $\boldsymbol{M}_X \triangleq \hat{\mu}_{X|\mho} \cdot \hat{\mu}^T_{X|\mho}$. Similarly, we define $\tilde{\boldsymbol{M}}_Y$, $\boldsymbol{M}_Y$ and $\hat{\mu}_{Y|\mho}$. According to (Huang et al., 2020), we have

$$\hat{\mu}_{X|\mho} \triangleq \phi(\mho)(\mathcal{C}_{\mho\mho} + \gamma I)^{-1}\mathcal{C}_{\mho X}, \tag{20}$$

where $\hat{\mu}_{X|\mho} \triangleq \phi(\mho)(\mathcal{C}_{\mho\mho} + \gamma I)^{-1}\mathcal{C}_{\mho X}$, $\hat{\mu}_{X|\mho}, \phi(\mho) \in \mathbb{R}^{n\times h}$, γ is a small ridge parameter, ϕ represents the feature map, and $\mho$ is the surrogate variable indicating different domains or clients. Similarly, we define $\tilde{\boldsymbol{M}}_Y$, $\boldsymbol{M}_Y$ and $\hat{\mu}_{Y|\mho}$.

$$\hat{\mu}_{Y|\mho} \triangleq \phi(\mho)(\mathcal{C}_{\mho\mho} + \gamma I)^{-1}\mathcal{C}_{\mho Y}. \tag{21}$$

Moreover, $\tilde{\boldsymbol{M}}_{\underline{Y}X} \triangleq \boldsymbol{H}\boldsymbol{M}_{\underline{Y}X}\boldsymbol{H}$, $\boldsymbol{M}_{\underline{Y}X} \triangleq \hat{\mu}_{\underline{Y}X|\mho} \cdot \hat{\mu}^T_{\underline{Y}X|\mho}$. Similarly, we define $\tilde{\boldsymbol{M}}_{\underline{X}Y}$, $\boldsymbol{M}_{\underline{X}Y}$ and $\hat{\mu}_{\underline{X}Y}$.

$$\begin{aligned}\hat{\mu}_{\underline{Y}X|\mho} &\triangleq \phi(\mho)(\mathcal{C}_{\mho\mho} + \gamma I)^{-1}\mathcal{C}_{\mho,(Y,X)}\\ \hat{\mu}_{\underline{X}Y|\mho} &\triangleq \phi(\mho)(\mathcal{C}_{\mho\mho} + \gamma I)^{-1}\mathcal{C}_{\mho,(X,Y)},\end{aligned} \tag{22}$$

Eq. 19 as formulated above is helpful to further derive Theorem 5 in our main paper.

A3 PROOFS

Here, we provide the proofs of the theorems and lemmas, including Lemma 3, Theorem 4, Theorem 5, Theorem 6, Lemma 7, and Theorem 8 in our main paper.

A3.1 PROOF OF LEMMA 3

Proof: We define the covariance matrix in the null hypothesis as $\mathcal{C}_{\ddot{X}Y|Z} = \frac{1}{n}\sum_{i=1}^{n}[(\ddot{A}_i - \mathbb{E}(\ddot{A}|Z))^T(B_i - \mathbb{E}(B|Z))]$ which corresponds to the partial cross-covariance matrix with n samples, $\mathcal{C}_{\ddot{X}Y|Z}\in\mathbb{R}^{h\times h}$, $\ddot{A}=f(\ddot{X})\in\mathbb{R}^{n\times h}$, $B=g(Y)\in\mathbb{R}^{n\times h}$, $\{f^j(\ddot{X})|_{j=1}^h\}\in\mathcal{F}_{\ddot{X}}$, $\{g^j(Y)|_{j=1}^h\}\in\mathcal{F}_Y$. Notice that $\mathcal{F}_{\ddot{X}}$ and $\mathcal{F}_Y$ are function spaces. n and h denote the number of total samples of all clients and the number of hidden features or mapping functions, respectively.

Notice that $\mathbb{E}(\ddot{A}|Z)$ and $\mathbb{E}(B|Z)$ could be non-linear functions of Z which may be difficult to estimate. therefore, we would like to approximate them with linear functions. Let $q(Z)\in\mathbb{R}^{n\times h}$, $\{q^j(Z)|_{j=1}^h\}\in\mathcal{F}_Z$. We could estimate $\mathbb{E}(f^j|Z)$ with the ridge regression output $u_j^T q(Z)$ under the mild conditions given below.

Lemma 11. (Sutherland & Schneider, 2015) *Consider performing ridge regression of f^j on Z. Assume that (i) $\sum_{i=1}^n f_i^j = 0$, f^j is defined on the domain of $\ddot{X}$; (ii) the empirical kernel matrix of Z, denoted by $\mathcal{K}_Z$, only has finite entries (i.e., $\|\mathcal{K}_Z\|_\infty < \infty$); (iii) the range of Z is compact, $Z \subset \mathbb{R}^{d_Z}$. Then we have*

$$\mathbb{P}\left[|\hat{\mathbb{E}}(f^j|Z) - u_j^T q(Z)| \geq \epsilon\right] \leq \frac{c_0}{\epsilon^2}e^{-h\epsilon^2 c_1}, \tag{23}$$

where $\hat{\mathbb{E}}(f^j|Z)$ is the estimate of $\mathbb{E}(f^j|Z)$ by ridge regression, c_0 and c_1 are both constants that do not depend on the sample size n or the number of hidden dimensions or mapping functions h.

The exponential rate with respect to h in the above lemma suggests we can approximate the output of ridge regression with a small number of hidden features. Moreover, we could similarly estimate $\mathbb{E}(g^j|Z)$ with $v_j^T q(Z)$, because we could guarantee that $\mathbb{P}\left[|\hat{\mathbb{E}}(g^j|Z) - v_j^T q(Z)| \geq \epsilon\right] \to 0$ for any fixed $\epsilon > 0$ at an exponential rate with respect to h.

Similar to the L^2 spaces in condition (ii) of Lemma 10, we can consider the following condition to approximate conditional independence:

$$\begin{aligned}&\mathbb{E}(\tilde{f}\tilde{g}) = 0, \forall \tilde{f} \in \tilde{\mathcal{F}}_{\ddot{X}|Z} \text{ and } \forall \tilde{g} \in \tilde{\mathcal{F}}_{Y|Z}, \text{ where}\\ &\tilde{\mathcal{F}}_{\ddot{X}|Z} = \{\tilde{f} \mid \tilde{f}^j = f^j - \mathbb{E}(f^j|Z), f^j \in \mathcal{F}_{\ddot{X}}\},\\ &\tilde{\mathcal{F}}_{Y|Z} = \{\tilde{g} \mid \tilde{g}^j = g^j - \mathbb{E}(g^j|Z), g^j \in \mathcal{F}_Y\}.\end{aligned} \tag{24}$$

According to Eq. 23, we could estimate $\mathbb{E}(f^j|Z)$ and $\mathbb{E}(g^j|Z)$ by $u_j^T q(Z)$ and $v_j^T q(Z)$, respectively. Thus, we can reformulate the function spaces as

$$\begin{aligned}\tilde{\mathcal{F}}_{\ddot{X}|Z} &= \{\tilde{f} \mid \tilde{f}^j = f^j - u_j^T q(Z), f^j \in \mathcal{F}_{\ddot{X}}\},\\ \tilde{\mathcal{F}}_{Y|Z} &= \{\tilde{g} \mid \tilde{g}^j = g^j - v_j^T q(Z), g^j \in \mathcal{F}_Y\}.\end{aligned} \tag{25}$$

Proof ends.

A3.2 PROOF OF THEOREM 4

Proof: Assume that there are n i.i.d. samples for X, Y, Z. Let $\tilde{\boldsymbol{K}}_{\ddot{X}|Z}$ be the centralized kernel matrix, given by $\tilde{\boldsymbol{K}}_{\ddot{X}|Z} \triangleq \tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{\ddot{X}|Z} = \boldsymbol{H}R_{\ddot{X}|Z}R^T_{\ddot{X}|Z}\boldsymbol{H}$, where $R_{\ddot{X}|Z} \triangleq \tilde{f}(\ddot{X}) = f(\ddot{X}) - u^T q(Z)$ which can be seen as the residual after ridge regression. Similarly, We could define $\tilde{\boldsymbol{K}}_{Y|Z} \triangleq \tilde{R}_{Y|Z}\tilde{R}^T_{Y|Z} = \boldsymbol{H}R_{Y|Z}R^T_{Y|Z}\boldsymbol{H}$ and $R_{Y|Z} \triangleq \tilde{g}(Y) = g(Y) - v^T q(Z)$. Accordingly, we let $\tilde{\boldsymbol{K}}_{\ddot{X}Y|Z} \triangleq \tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{Y|Z} = \boldsymbol{H}R_{\ddot{X}|Z}R^T_{Y|Z}\boldsymbol{H}$. We set the test statistic as $\mathcal{T}_{CI} = n\|\mathcal{C}_{\ddot{X}Y|Z}\|^2_F$, where $\mathcal{C}_{\ddot{X}Y|Z} \triangleq \tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{Y|Z} = \frac{1}{n}R^T_{\ddot{X}|Z}\boldsymbol{H}\boldsymbol{H}R_{Y|Z}$.

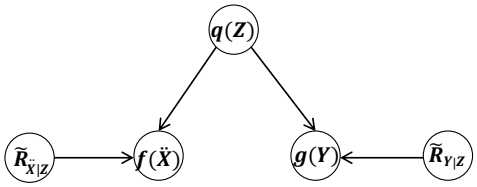


Figure A1: Given that $X \perp\!\!\!\perp Y|Z$, we could introduce the independence between $R_{\ddot{X}|Z}$ and $R_{Y|Z}$.

Let $\lambda_{\ddot{X}|Z}$ and $\lambda_{Y|Z}$ be the eigenvalues of $\tilde{\boldsymbol{K}}_{\ddot{X}|Z}$ and $\tilde{\boldsymbol{K}}_{Y|Z}$, respectively. Furthermore, we define the EVD decomposition $\tilde{\boldsymbol{K}}_{\ddot{X}|Z} = \boldsymbol{V}_{\ddot{X}|Z}\boldsymbol{\Lambda}_{\ddot{X}|Z}\boldsymbol{V}^T_{\ddot{X}|Z}$, where $\boldsymbol{\Lambda}_{\ddot{X}|Z}$ is the diagonal matrix containing non-negative eigenvalues $\lambda_{\ddot{X}|Z,i}$. Similarly, we define $\tilde{\boldsymbol{K}}_{Y|Z} = \boldsymbol{V}_{Y|Z}\boldsymbol{\Lambda}_{Y|Z}\boldsymbol{V}^T_{Y|Z}$ with eigenvalues $\lambda_{Y|Z,i}$. Let $\boldsymbol{\psi}_{\ddot{X}|Z} = [\psi_{\ddot{X}|Z,1}, \psi_{\ddot{X}|Z,2}, \ldots, \psi_{\ddot{X}|Z,n}] \triangleq \boldsymbol{V}_{Y|Z}\boldsymbol{\Lambda}^{1/2}_{Y|Z}$ and $\boldsymbol{\phi}_{Y|Z} = [\phi_{Y|Z,1}, \phi_{Y|Z,2}, \ldots, \phi_{Y|Z,n}] \triangleq \boldsymbol{V}_{Y|Z}\boldsymbol{\Lambda}^{1/2}_{Y|Z}$.

On the other hand, consider eigenvalues $\lambda^*_{\ddot{X}|Z,i}$ and eigenfunctions $u_{\ddot{X}|Z,i}$ of the kernel $k_{\ddot{X}|Z}$ w.r.t. the probablity measure with the density $\mathbb{P}(\ddot{x})$, i.e., $\lambda^*_{\ddot{X}|Z,i}$ and $u_{\ddot{X}|Z,i}$ satisfy $\int k_{\ddot{X}|Z}(\ddot{x}, \ddot{x}') \cdot u_{\ddot{X}|Z,i}(\ddot{x}) \cdot \mathbb{P}(\ddot{x})d\ddot{x} = \lambda^*_{\ddot{X}|Z,i} \cdot u_{\ddot{X}|Z,i}(\ddot{x}')$, where we assume that $u_{\ddot{X}|Z,i}$ have unit variance, i.e., $\mathbb{E}[u^2_{\ddot{X}|Z,i}(\ddot{X})] = 1$. Similarly, we define $k_{Y|Z}$, $\lambda^*_{Y|Z,i}$, and $u^*_{Y|Z,i}$. Let $\{\alpha_1, \ldots, \alpha_L\}$ denote i.i.d. standard Gaussian variables, and thus $\{\alpha^2_1, \ldots, \alpha^2_L\}$ denote i.i.d. χ^2_1 variables.

Lemma 12 (Kernel-based Conditional Independence Test (Zhang et al., 2012)). *Under the null hypothesis that X and Y are conditional independent given Z, we have that the test statistic $\mathcal{T}_{CI} \triangleq \frac{1}{n}\operatorname{tr}(\tilde{\boldsymbol{K}}_{\ddot{X}|Z}\tilde{\boldsymbol{K}}_{Y|Z})$ have the same asymptotic distribution as $\hat{\mathcal{T}}_{CI} \triangleq \frac{1}{n}\sum_{k=1}^{n^2}\tilde{\lambda}_k \cdot \alpha^2_k$, where $\tilde{\lambda}_k$ are eigenvalues of $\boldsymbol{w}\boldsymbol{w}^T$, $\boldsymbol{w} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]$, with the vector $\boldsymbol{w}_t$ obtained by stacking $\boldsymbol{M}_t = [\psi_{\ddot{X}|Z,1}(\ddot{X}_t), \psi_{\ddot{X}|Z,2}(\ddot{X}_t), \ldots, \psi_{\ddot{X}|Z,n}(\ddot{X}_t)]^T \cdot [\phi_{Y|Z,1}(Y_t), \phi_{Y|Z,2}(Y_t), \ldots, \phi_{Y|Z,n}(Y_t)]$.*

In the above lemma, their test statistic is equivalent to ours, due to the fact that

$$\begin{aligned}
\frac{1}{n}\operatorname{tr}(\tilde{\boldsymbol{K}}_{\ddot{X}|Z}\tilde{\boldsymbol{K}}_{Y|Z}) &= \frac{1}{n}\operatorname{tr}(\tilde{R}_{\ddot{X}|Z}(\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{Y|Z}\tilde{R}^T_{Y|Z})) \\
&= \frac{1}{n}\operatorname{tr}((\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{Y|Z}\tilde{R}^T_{Y|Z})\tilde{R}_{\ddot{X}|Z}) \\
&= \frac{1}{n}\|\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{Y|Z}\|^2_F \\
&= \frac{1}{n}\|n\mathcal{C}_{\ddot{X}Y|Z}\|^2_F \\
&= n\|\mathcal{C}_{\ddot{X}Y|Z}\|^2_F.
\end{aligned} \tag{26}$$

However, their asymptotic distribution is different from ours. Based on their asymptotic distribution, we could go further. The first two rows of Eq. 26 hold true because of the commutative property of trace, namely, $\operatorname{tr}(AB) = \operatorname{BA}$, refer to Lemma 6 for more details. According to the formulation of $\tilde{R}_{\ddot{X}|Z}$ and $\tilde{R}_{Y|Z}$, we have

$$\begin{cases} f(\ddot{X}) = u^T q(Z) + R_{\ddot{X}|Z} \\ g(Y) = v^T q(Z) + R_{Y|Z}. \end{cases} \tag{27}$$

Based on the above formulations, we could easily draw the causal graph as shown in Fig. A1. In particular, considering that X and Y are conditionally independent given Z, we could further determine that $R_{\ddot{X}|Z}$ and $R_{Y|Z}$ are independent, namely, we have

$$X \perp\!\!\!\perp Y|Z \iff R_{\ddot{X}|Z} \perp\!\!\!\perp R_{Y|Z}. \tag{28}$$

As $f(\ddot{X})$ and $g(Y)$ are uncorrelated, then $\mathbb{E}(\boldsymbol{w}_t) = 0$. Furthermore, the covariance is $\boldsymbol{\Sigma} = \mathbb{C}\mathrm{ov}(\boldsymbol{w}_t) = \mathbb{E}(\boldsymbol{w_t}\boldsymbol{w_t^T})$, where $\boldsymbol{w}$ is defined in the same way as in Lemma 12. If $R_{\ddot{X}|Z} \perp\!\!\!\perp R_{Y|Z}$, for $k \neq i$ or $l \neq j$, we denote the non-diagonal (ND) entries of $\boldsymbol{\Sigma}$ as e_{ND}, where

$$\begin{aligned} e_{ND} &= \mathbb{E}[\sqrt{\lambda^*_{\ddot{X}|Z,i}\lambda^*_{Y|Z,j}\lambda^*_{\ddot{X}|Z,k}\lambda^*_{Y|Z,l}}u_{\ddot{X}|Z,i}u_{Y|Z,j}u_{\ddot{X}|Z,k}u_{Y|Z,l}] \\ &= \sqrt{\lambda^*_{\ddot{X}|Z,i}\lambda^*_{Y|Z,j}\lambda^*_{\ddot{X}|Z,k}\lambda^*_{Y|Z,l}}\mathbb{E}[u_{\ddot{X}|Z,i}u_{\ddot{X}|Z,k}]\mathbb{E}[u_{Y|Z,j}u_{Y|Z,l}] \\ &= 0. \end{aligned} \tag{29}$$

We then denote the diagonal entries of $\boldsymbol{\Sigma}$ as e_D, where

$$\begin{aligned} e_D &= \lambda^*_{\ddot{X}|Z,i}\lambda^*_{Y|Z,j}\mathbb{E}[u^2_{\ddot{X}|Z,i}]\mathbb{E}[u^2_{Y|Z,j}] \\ &= \lambda^*_{\ddot{X}|Z,i}\lambda^*_{Y|Z,j}, \end{aligned} \tag{30}$$

which are eigenvalues of Σ. According to (Zhang et al., 2012), $\frac{1}{n}\lambda_{\ddot{X}|Z,i}$ converge in probability $\lambda^*_{\ddot{X}|Z}$. Substituting all the results into the asymptotic distribution in Lemma 12, we can get the updated asymptotic distribution

$$\hat{\mathcal{T}}_{CI} \triangleq \frac{1}{n^2}\sum_{i,j=1}^{L}\lambda_{\ddot{X}|Z,i}\lambda_{Y|Z,j}\alpha^2_{ij} \quad \text{as} \quad L = n \to \infty. \tag{31}$$

Consequently, $\mathcal{T}_{CI}$ and $\hat{\mathcal{T}}_{CI}$ have the same asymptotic distribution. Proof ends.

A3.3 Proof of Theorem 5

Proof: First of all, since α^2_{ij} follow the χ^2 distribution with one degree of freedom, thus we have $\mathbb{E}(\alpha^2_{ij}) = 1$ and $\mathbb{V}ar(\alpha^2_{ij}) = 2$. According to the asymptotic distribution in Theorem 4 and the derivation of Lemma 7, we have

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{T}}_{CI}|\mathcal{D}) &= \frac{1}{n^2}\sum_{i,j}\lambda_{\ddot{X}|Z,i}\lambda_{Y|Z,j} \\ &= \frac{1}{n^2}\sum_{i}\lambda_{\ddot{X}|Z,i}\sum_{j}\lambda_{Y|Z,j} \\ &= \frac{1}{n^2}\,\mathrm{tr}(\tilde{\boldsymbol{K}}_{\ddot{X}|Z})\,\mathrm{tr}(\tilde{\boldsymbol{K}}_{Y|Z}) \\ &= \frac{1}{n^2}\,\mathrm{tr}(\tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{\ddot{X}|Z})\,\mathrm{tr}(\tilde{R}_{Y|Z}\tilde{R}^T_{Y|Z}) \\ &= \frac{1}{n^2}\,\mathrm{tr}(n\cdot\mathcal{C}_{\ddot{X}|Z})\,\mathrm{tr}(n\cdot\mathcal{C}_{Y|Z}) \\ &= \mathrm{tr}(\mathcal{C}_{\ddot{X}|Z})\,\mathrm{tr}(\mathcal{C}_{Y|Z}), \end{aligned} \tag{32}$$

where $\tilde{R}_{\ddot{X}|Z}$ and $\tilde{R}_{Y|Z}$ are defined in the proof of Theorem 3 above. Therefore, $\mathbb{E}(\hat{\mathcal{T}}_{CI}|\mathcal{D}) = \mathrm{tr}(\mathcal{C}_{\ddot{X}|Z})\,\mathrm{tr}(\mathcal{C}_{Y|Z})$.

Furthermore, α_{ij}^2 are independent variables across i and j, and notice that $\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{\ddot{X}|Z}) = \sum_i \lambda^2_{\ddot{X}|Z,i}$, and similarly $\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{Y|Z}) = \sum_i \lambda^2_{Y|Z,i}$. Based on the asymptotic distribution in Theorem 4, we have

$$\begin{aligned}
\mathbb{V}ar(\hat{\mathcal{T}}_{CI}|\mathcal{D}) &= \frac{1}{n^4}\sum_{i,j}\lambda^2_{\ddot{X}|Z,i}\lambda^2_{Y|Z,j}\mathbb{V}ar(\alpha^2_{ij}) \\
&= \frac{2}{n^4}\sum_{i}\lambda^2_{\ddot{X}|Z,i}\sum_{j}\lambda^2_{Y|Z,j} \\
&= \frac{2}{n^4}\,\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{\ddot{X}|Z})\,\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{Y|Z}).
\end{aligned} \tag{33}$$

Additionally, according to the similar rule as in Eq. 26, we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{\ddot{X}|Z}) &= \mathrm{tr}(\tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{\ddot{X}|Z}) \\
&= \mathrm{tr}(\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{\ddot{X}|Z}\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{\ddot{X}|Z}) \\
&= \|\tilde{R}^T_{\ddot{X}|Z}\tilde{R}_{\ddot{X}|Z}\|^2_F \\
&= \|n\cdot\mathcal{C}_{\ddot{X}|Z}\|^2_F \\
&= n^2\|\mathcal{C}_{\ddot{X}|Z}\|^2_F.
\end{aligned} \tag{34}$$

Similarly, we have $\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{Y|Z}) = n^2\|\mathcal{C}_{Y|Z}\|^2_F$. Substituting the results into the above formulation about variance, we have $\frac{2}{n^4}\,\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{\ddot{X}|Z})\,\mathrm{tr}(\tilde{\boldsymbol{K}}^2_{Y|Z}) = \frac{2}{n^4}\cdot n^2\|\mathcal{C}_{\ddot{X}|Z}\|^2_F \cdot n^2\|\mathcal{C}_{Y|Z}\|^2_F$. Thus, $\mathbb{V}ar(\hat{\mathcal{T}}_{CI}|\mathcal{D}) = 2\cdot\|\mathcal{C}_{\ddot{X}|Z}\|^2_F\cdot\|\mathcal{C}_{Y|Z}\|^2_F$. Proof ends.

A3.4 Proof of Theorem 6

Proof: According to the above-mentioned formulations, we have $\tilde{\boldsymbol{M}}_X \triangleq \boldsymbol{H}\boldsymbol{M}_X\boldsymbol{H} = \tilde{\hat{\mu}}_{X|\mho}\cdot\tilde{\hat{\mu}}^T_{X|\mho}$, $\tilde{\hat{\mu}}_{X|\mho} \triangleq \boldsymbol{H}\cdot\hat{\mu}_{X|\mho}$. Based on the rules of estimating covariance matrix from kernel matrix in Lemma 6, we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{M}}_X) &= \mathrm{tr}(\tilde{\hat{\mu}}_{X|\mho}\cdot\tilde{\hat{\mu}}^T_{X|\mho}) \\
&= \mathrm{tr}(\tilde{\hat{\mu}}^T_{X|\mho}\cdot\tilde{\hat{\mu}}_{X|\mho}) && (35)\\
&= \mathrm{tr}((\boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho X})^T(\boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho X})) && (36)\\
&= \mathrm{tr}(\mathcal{C}_{X\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\phi(\mho)^T\boldsymbol{H}\cdot\boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho X})) \\
&= \frac{1}{n}\,\mathrm{tr}(\mathcal{C}_{X\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho X}) && (37)\\
&= \frac{1}{n}\,\mathrm{tr}(\mathcal{C}^*_X). && (38)
\end{aligned}$$

Eq. 35 is obtained due to the trace property of the product of the matrices, as shown in Lemma 6. Eq. 36 is substituting from Eq. 20. Here we use Eq. 38 for simple notation. We can see that it can be represented with some combinations of different covariance matrices. Similarly, we have

$$\mathrm{tr}(\tilde{\boldsymbol{M}}_Y) = \frac{1}{n}\,\mathrm{tr}(\mathcal{C}_{Y\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho Y}) = \frac{1}{n}\,\mathrm{tr}(\mathcal{C}^*_Y). \tag{39}$$

Regarding the centralized Gram matrices for joint distribution, similarly we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{M}}_{\underline{Y}X}) &= \frac{1}{n}\,\mathrm{tr}(\mathcal{C}_{(Y,X),\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(Y,X)}) = \frac{1}{n}\,\mathrm{tr}(\mathcal{C}^*_{\tilde{Y}}), \\
\mathrm{tr}(\tilde{\boldsymbol{M}}_{\underline{X}Y}) &= \frac{1}{n}\,\mathrm{tr}(\mathcal{C}_{(X,Y),\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(X,Y)}) = \frac{1}{n}\,\mathrm{tr}(\mathcal{C}^*_{\tilde{X}}),
\end{aligned} \tag{40}$$

where $\mathrm{tr}(\tilde{\boldsymbol{M}}_{\underline{Y}X}) = \mathrm{tr}(\tilde{\boldsymbol{M}}_{\underline{X}Y})$. Furthermore, based on Lemma 6 and Eq. 22, we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{M}}_X \tilde{\boldsymbol{M}}_{\underline{Y}X}) &= \mathrm{tr}(\tilde{\hat{\mu}}_{X|\mho}\tilde{\hat{\mu}}^T_{X|\mho} \cdot \tilde{\hat{\mu}}_{\underline{Y}X|\mho}\tilde{\hat{\mu}}^T_{\underline{Y}X|\mho}) \\
&= \mathrm{tr}(\tilde{\hat{\mu}}^T_{X|\mho}\tilde{\hat{\mu}}_{\underline{Y}X|\mho}\tilde{\hat{\mu}}^T_{\underline{Y}X|\mho} \cdot \tilde{\hat{\mu}}_{X|\mho}) \qquad (41) \\
&= \|\tilde{\hat{\mu}}^T_{X|\mho}\tilde{\hat{\mu}}_{\underline{Y}X|\mho}\|_F^2 \\
&= \|(\boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho X})^T(\boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(Y,X)})\|_F^2 \qquad (42) \\
&= \|\mathcal{C}_{X\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\phi(\mho)^T\boldsymbol{H} \cdot \boldsymbol{H}\phi(\mho)(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(Y,X)}\|_F^2 \\
&= \|\frac{1}{n}\mathcal{C}_{X\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(Y,X)}\|_F^2 \qquad (43) \\
&= \frac{1}{n^2}\|\mathcal{C}^*_{X,\tilde{Y}}\|_F^2. \qquad (44)
\end{aligned}$$

Eq. 41 is obtained due to the trace property of the product of the matrices, as shown in Lemma 6. Eq. 41 is substituting from Eq. 20 and Eq. 22. Here we use Eq. 44 for simple notation. We can see that it can be represented with some combinations of different covariance matrices. Similarly, we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{M}}_Y \tilde{\boldsymbol{M}}_{\underline{X}Y}) &= \|\frac{1}{n}\mathcal{C}_{Y\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho\mho}(\mathcal{C}_{\mho\mho}+\gamma I)^{-1}\mathcal{C}_{\mho,(X,Y)}\|_F^2 \\
&= \frac{1}{n^2}\|\mathcal{C}^*_{Y,\tilde{X}}\|_F^2.
\end{aligned} \qquad (45)$$

Substituting the equations above into Eq. 19, we have

$$\hat{\triangle}_{X\to Y} = \frac{\|\mathcal{C}^*_{X,\tilde{Y}}\|_F^2}{\mathrm{tr}(\mathcal{C}^*_X)\cdot\mathrm{tr}(\mathcal{C}^*_{\tilde{Y}})}, \quad \hat{\triangle}_{Y\to X} = \frac{\|\mathcal{C}^*_{Y,\tilde{X}}\|_F^2}{\mathrm{tr}(\mathcal{C}^*_Y)\cdot\mathrm{tr}(\mathcal{C}^*_{\tilde{X}})}. \qquad (46)$$

Proof ends.

A3.5 Proof of Lemma 7

Proof: First of all, we incorporate random Fourier features to approximate the kernels, because they have shown competitive performances to approximate the continuous shift-invariant kernels.

Lemma 13 (Random Features (Rahimi & Recht, 2007)). *For a continuous shift-invariant kernel $\mathcal{K}(x,y)$ on $\mathbb{R}$, we have:*

$$\mathcal{K}(x,y) = \int_{\mathbb{R}} p(w)e^{jw(x-y)}dw = \mathbb{E}_w[\zeta_w(x)\zeta_w(y)], \qquad (47)$$

where $\zeta_w(x)\zeta_w(y)$ is an unbiased estimate of $\mathcal{K}(x,y)$ when w is drawn from $p(w)$.

Since both the probability distribution $p(w)$ and the kernel entry $\mathcal{K}(x,y)$ are real, the integral in Eq. 47 converges when the complex exponentials are replaced with cosines. Therefore, we may get a real-values mapping by:

$$\begin{aligned}
\mathcal{K}(x,y) &\approx \phi_w(x)^T\phi_w(y), \\
\phi_w(x) &\triangleq \sqrt{\frac{2}{h}}[\cos(w_1x+b_1),...,\cos(w_hx+b_h)]^T, \\
\phi_w(y) &\triangleq \sqrt{\frac{2}{h}}[\cos(w_1y+b_1),...,\cos(w_hy+b_h)]^T,
\end{aligned} \qquad (48)$$

where w is drawn from $p(w)$ and b is drawn uniformly from $[0, 2\pi]$. $x, y, w, b \in \mathbb{R}$, and the randomized feature map $\phi_w : \mathbb{R} \to \mathbb{R}^h$. The precise form of $p(w)$ relies on the type of the shift-invariant kernel we would like to approximate. Here in this paper, we choose to approximate Gaussian kernel as one of the characteristic kernels, and thus set the probability distribution $p(w)$ to the Gaussian one. Based on Eq. 48, we have

$$\mathrm{tr}(\tilde{\boldsymbol{K}}_{\boldsymbol{x},\boldsymbol{y}}) \approx \mathrm{tr}(\tilde{\phi}_w(\boldsymbol{x})\tilde{\phi}_w(\boldsymbol{y})^T), \qquad (49)$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\tilde{\boldsymbol{K}}_{\boldsymbol{x},\boldsymbol{y}} \in \mathbb{R}^{n\times n}$, $\tilde{\phi}_w(\boldsymbol{x}) \in \mathbb{R}^{n\times h}$ is the centralized random feature, $\tilde{\phi}_w(\boldsymbol{x}) = \boldsymbol{H}\phi_w(\boldsymbol{x})$. Furthermore, benefiting from the commutative property of the trace of the product of two matrices, we have

$$\operatorname{tr}(\tilde{\phi}_w(\boldsymbol{x})\tilde{\phi}_w(\boldsymbol{y})^T) = \operatorname{tr}(\tilde{\phi}_w(\boldsymbol{y})^T\tilde{\phi}_w(\boldsymbol{x})), \tag{50}$$

Since each random feature is centralized, meaning the zero mean for each feature, therefore, we have:

$$\operatorname{tr}(\tilde{\phi}_w(\boldsymbol{y})^T\tilde{\phi}_w(\boldsymbol{x})) = \operatorname{tr}(\frac{1}{n}\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{n}\operatorname{tr}(\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}}), \tag{51}$$

where $\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}}$ is the covariance matrix for variable x and y, $\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}} \in \mathbb{R}^{h\times h}$, h is the number of hidden features.

For the second formulation, we have

$$\begin{aligned}
\operatorname{tr}(\tilde{\boldsymbol{K}}_{\boldsymbol{x}}\tilde{\boldsymbol{K}}_{\boldsymbol{y}}) &= \operatorname{tr}[\tilde{\phi}_w(\boldsymbol{x})\tilde{\phi}_w(\boldsymbol{x})^T\tilde{\phi}_w(\boldsymbol{y})\tilde{\phi}_w(\boldsymbol{y})^T] \\
&= \operatorname{tr}[\tilde{\phi}_w(\boldsymbol{x})(\tilde{\phi}_w(\boldsymbol{x})^T\tilde{\phi}_w(\boldsymbol{y})\tilde{\phi}_w(\boldsymbol{y})^T)] \\
&= \operatorname{tr}[(\tilde{\phi}_w(\boldsymbol{x})^T\tilde{\phi}_w(\boldsymbol{y})\tilde{\phi}_w(\boldsymbol{y})^T)\tilde{\phi}_w(\boldsymbol{x})] \\
&= \operatorname{tr}[\tilde{\phi}_w(\boldsymbol{x})^T\tilde{\phi}_w(\boldsymbol{y})\tilde{\phi}_w(\boldsymbol{y})^T\tilde{\phi}_w(\boldsymbol{x})] \\
&= \|\tilde{\phi}_w(\boldsymbol{x})^T\tilde{\phi}_w(\boldsymbol{y})\|_F^2 \\
&= \|n\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}}\|_F^2 \\
&= n^2\|\mathcal{C}_{\boldsymbol{x},\boldsymbol{y}}\|_F^2.
\end{aligned} \tag{52}$$

Together with Eq. 49, Eq. 50, Eq. 51 and Eq. 52 formulated above, we could prove the Lemma 7 in the main paper. Proof ends.

A3.6 Proof of Theorem 8

Proof. The summary statistics contain two parts: total sample size n and covariance tensor $\mathcal{C}_{\mathcal{T}} \in \mathbb{R}^{d'\times d'\times h\times h}$. Let $\mathcal{C}_{\mathcal{T}}^{ij} \in \mathbb{R}^{h\times h}$ be the (i,j)-th entry of the covariance tensor, which denotes the covariance matrix of the i-th and the j-th variable.

With the summary statistics as a proxy, we can substitute the raw data at each client. During the procedures of causal discovery, the needed statistics include $\mathcal{T}_{CI}$ in Theorem 4, $\mathbb{E}(\hat{\mathcal{T}}_{CI}|\mathcal{D})$ and $\mathbb{V}ar(\hat{\mathcal{T}}_{CI}|\mathcal{D})$ in Theorem 5, and $\hat{\triangle}_{X\to Y}$ and $\hat{\triangle}_{Y\to X}$ in Theorem 6.

1) Based on the Eq. (7) in the main paper, we have

$$\begin{aligned}
\mathcal{C}_{\ddot{X}Y|Z} &= \mathcal{C}_{\ddot{X}Y} - \mathcal{C}_{\ddot{X}Z}(\mathcal{C}_{ZZ} + \gamma I)^{-1}\mathcal{C}_{ZY} \\
&= \mathcal{C}_{(X,Z),Y} - \mathcal{C}_{(X,Z),Z}(\mathcal{C}_{ZZ} + \gamma I)^{-1}\mathcal{C}_{ZY} \\
&= (\mathcal{C}_{XY} + \mathcal{C}_{ZY}) - (\mathcal{C}_{XZ} + \mathcal{C}_{ZZ})(\mathcal{C}_{ZZ} + \gamma I)^{-1}\mathcal{C}_{ZY}
\end{aligned} \tag{53}$$

In this paper, we consider the scenarios where X and Y are single variables, and Z may be a single variable, a set of variables, or empty. Assuming that Z contains L variables. We have

$$\mathcal{C}_{ZY} = \sum_{i=1}^{L}\mathcal{C}_{Z_iY}, \;\; \mathcal{C}_{XZ} = \sum_{i=1}^{L}\mathcal{C}_{XZ_i}, \;\; \mathcal{C}_{ZZ} = \sum_{i=1}^{L}\sum_{j=1}^{L}\mathcal{C}_{Z_iZ_j}, \tag{54}$$

where $\mathcal{C}_{XY}, \mathcal{C}_{Z_iY}, \mathcal{C}_{XZ_i}$, and $\mathcal{C}_{Z_iZ_j}$ are the entries of the covariance tensor $\mathcal{C}_{\mathcal{T}}$. According to Theorem 3, $\mathcal{T}_{CI} \triangleq n\|\mathcal{C}_{\ddot{X}Y|Z}\|_F^2$. Therefore, the summary statistics are sufficient to represent $\mathcal{T}_{CI}$.

2) Similar to Eq. 53, we have

$$\mathcal{C}_{\ddot{X}|Z} = (\mathcal{C}_{XX} + 2\mathcal{C}_{XZ} + \mathcal{C}_{ZZ})(\mathcal{C}_{XZ} + \mathcal{C}_{ZZ})(\mathcal{C}_{ZZ} + \gamma I)^{-1}(\mathcal{C}_{XZ} + \mathcal{C}_{ZZ}) \tag{55}$$

$$\mathcal{C}_{Y|Z} = \mathcal{C}_{YY} - \mathcal{C}_{YZ}(\mathcal{C}_{ZZ} + \gamma I)^{-1}\mathcal{C}_{ZY}. \tag{56}$$

Substituting Eq. 54 into Eq. 55 and Eq. 56, we can also conclude that the covariance tensor is sufficient to represent $\mathcal{C}_{\ddot{X}|Z}$ and $\mathcal{C}_{Y|Z}$. In other words, the summary statistics are sufficient to represent $\mathbb{E}(\hat{\mathcal{T}}_{CI}|\mathcal{D})$ and $\mathbb{V}ar(\hat{\mathcal{T}}_{CI}|\mathcal{D})$.

3) As shown in section A3.3, we have

$$\hat{\triangle}_{X\to Y} = \frac{\|\mathcal{C}^*_{X,\tilde{Y}}\|_F^2}{\mathrm{tr}(\mathcal{C}^*_X)\cdot\mathrm{tr}(\mathcal{C}^*_{\tilde{Y}})}, \quad \hat{\triangle}_{Y\to X} = \frac{\|\mathcal{C}^*_{Y,\tilde{X}}\|_F^2}{\mathrm{tr}(\mathcal{C}^*_Y)\cdot\mathrm{tr}(\mathcal{C}^*_{\tilde{X}})}, \tag{57}$$

where each components can be represented as some combinations of covariance matrices, as shown in Eq. 37, Eq. 39, Eq. 40, Eq. 43, and Eq. 45. Therefore, the summary statistics are sufficient to represent $\hat{\triangle}_{X\to Y}$ and $\hat{\triangle}_{Y\to X}$.

4) To sum up, we could conclude that: The summary statistics, consisting of total sample size n and covariance tensor $\mathcal{C}_{\mathcal{T}}$, are sufficient to represent all the statistics needed for federated causal discovery.

Proof ends.

A4 Details about Federated Unconditional Independence Test

Here, we provide more details about the federated unconditional independence test (FUIT), where the conditioning set Z is empty. Generally, this method follows similar theorems for federated conditional independent test (FCIT).

A4.1 Null Hypothesis

Consider the null and alternative hypothesis

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y, \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y. \tag{58}$$

Similar to FCIT, we consider the squared Frobenius norm of the empirical covariance matrix as an approximation, given as

$$\mathcal{H}_0 : \|\mathcal{C}_{\ddot{X}Y}\|_F^2 = 0, \quad \mathcal{H}_1 : \|\mathcal{C}_{\ddot{X}Y}\|_F^2 > 0. \tag{59}$$

In this unconditional case, we set the test statistics as $\mathcal{T}_{UI} \triangleq n\|\mathcal{C}_{\ddot{X}Y}\|_F^2$, and give the following theorem.

Theorem 14 (Federated Unconditional Independent Test)**.** *Under the null hypothesis $\mathcal{H}_0$ (X and Y are independent), the test statistic*

$$\mathcal{T}_{UI} \triangleq n\|\mathcal{C}_{XY}\|_F^2, \tag{60}$$

has the asymptotic distribution

$$\hat{\mathcal{T}}_{UI} \triangleq \frac{1}{n^2}\sum_{i,j=1}^{L}\lambda_{X,i}\lambda_{Y,j}\alpha_{ij}^2,$$

where λ_X and λ_Y are the eigenvalues of $\tilde{\boldsymbol{K}}_X$ and $\tilde{\boldsymbol{K}}_Y$, respectively. Here, the proof is similar to the proof of Theorem 3, thus we refer the readers to section A3.2 for more details.

A4.2 Null Distribution Approximation

We also approximate the null distribution with a two-parameter Gamma distribution, which is related to the mean and variance. Under the hypothesis $\mathcal{H}_0$ and given the sample $\mathcal{D}$, the distribution of $\hat{\mathcal{T}}_{CI}$ can be approximated by the $\Gamma(\kappa,\theta)$ distribution. Here we provide the theorem for null distribution approximation.

Theorem 15 (Null Distribution Approximation)**.** *Under the null hypothesis $\mathcal{H}_0$ (X and Y are independent), we have*

$$\begin{aligned}\mathbb{E}(\hat{\mathcal{T}}_{UI}|\mathcal{D}) &= \mathrm{tr}(\mathcal{C}_X)\cdot\mathrm{tr}(\mathcal{C}_Y),\\ \mathbb{V}ar(\hat{\mathcal{T}}_{UI}|\mathcal{D}) &= 2\|\mathcal{C}_X\|_F^2\cdot\|\mathcal{C}_Y\|_F^2,\end{aligned} \tag{61}$$

Here, the proof is similar to the proof of Theorem 4, thus we refer the readers to section A3.3 for more details.

A5 Details about Skeleton Discovery and Direction Determination

In this section, we will introduce how we do the skeleton discovery and direction determination during the process of federated causal discovery. All those steps are conducted on the server side. Our steps are similar to the previous method, such as CD-NOD (Huang et al., 2020), the core difference are that we develop and utilize our proposed federated conditional independent test (FCIT) and federated independent change principle (FICP).

A5.1 Skeleton Discovery.

We first conduct skeleton discovery on the augmented graph. The extra surrogate variable is introduced in order to deal with the data heterogeneity across different clients.

Lemma 16. *Given the Assumptions 1, 2 and 3 in the main paper, for each $V_i \in \boldsymbol{V}$, V_i and $\mho$ are not adjacent in the graph if and only if they are independent conditional on some subset of $\{V_j | j \neq i\}$.*

Proof. If V_i's causal module is invariant, which means that $\mathbb{P}(V_i|PA_i)$ remains the same for every value of $\mho$, then $V_i \perp\!\!\!\perp \mho | PA_i$. Thus, if V_i and $\mho$ are not independent conditional on any subset of other variables, V_i's module changes with $\mho$, which is represented by an edge between V_i and $\mho$. Conversely, we assume that if V_i's module changes, which entails that V_i and $\mho$ are not independent given PA_i, then V_i and $\mho$ are not independent given any other subset of $\boldsymbol{V} \backslash \{V_i\}$. Proof ends.

Lemma 17. *Given the Assumptions 1, 2 and 3 in the main paper, for every $V_i, V_j \in \boldsymbol{V}$, V_i and V_j are not adjacent if and only if they are independent conditional on some subset of $\{V_l | l \neq i, l \neq j\} \cup \{\mho\}$.*

Proof. The "if" direction shown based on the faithfulness assumption on $\mathcal{G}_{aug}$ and the fact that $\{\psi_l(\mho)\}_{l=1}^L \cup \{\theta_i(\mho)\}_{i=1}^d$ is a deterministic function of $\mho$. The "only if" direction is proven by making use of the weak union property of conditional independence repeatedly, the fact that all $\{\psi_l(\mho)\}_{l=1}^L$ and $\{\theta_i(\mho)\}_{i=1}^d$ are deterministic function of $\mho$, the above three assumptions, and the properties of mutual information. Please refer to (Zhang et al., 2015) for more complete proof.

With the given three assumptions in the main paper, we can do skeleton discovery.

i) Augmented graph initialization. First of all, build a completely undirected graph on the extended variable set $\boldsymbol{V} \cup \{\mho\}$, where $\boldsymbol{V}$ denotes the observed variables and $\mho$ is surrogate variable.

ii) Changing module detection. For each edge $\mho - V_i$, conduct the federated conditional independence test or federated unconditional independent test. If they are conditionally independent or independent, remove the edge between them. Otherwise, keep the edge and orient $\mho \to V_i$.

iii) Skeleton discovery. Moreover, for each edge $V_i - V_j$, also conduct the federated independence test or federated unconditional independent test. If they are conditionally independent or independent, remove the edge between them.

In the procedures, how observed variables depend on surrogate variable $\mho$ is unknown and usually nonlinear, thus it is crucial to use a general and non-parametric conditional independent test method, which should also satisfy the federated learning constraints. Here, we utilize our proposed FCIT.

A5.2 Direction Determination.

After obtaining the skeleton, we can go on with the causal direction determination. By introducing the surrogate variable $\mho$, it does not only allow us to infer the skeleton, but also facilitate the direction determinations. For each variable V_i whose causal module is changing (i.e., $\mho - V_i$), in some ways we might determine the directions of every edge incident to V_i. Assume another variable V_j which is adjacent to V_i, then we can determine the directions via the following rules.

i) Direction determination with one changing module. When V_j's causal module is not changing, we can see $\mho - V_i - V_j$ forms an unshielded triple. For practice purposes, we can take the direction between $\mho$ and V_i as $\mho \to V_i$, since we let $\mho$ be the surrogate variable to indicate whether this causal module is changing or not. Then we can use the standard orientation rules (Spirtes et al., 2000) for unshielded triples to orient the edge between V_i and V_j. (1) If $\mho$ and V_i are independent conditional on some subset of $\{V_l | l \neq j\}$ which is excluding V_j, then the triple forms a V-structure, thus we have $\mho \to V_i \leftarrow V_j$. (2) If $\mho$ and V_i are independent conditional on some subset of $\{V_l | l \neq i\} \cup \{V_j\}$ which is including V_j, then we have $\mho \to V_i \to V_j$. In the procedure, we apply our proposed FCIT.

ii) Direction determination with two changing modules. When V_j's causal module is changing, we can see there is a special confounder $\mho$ between $V_i - V_j$. First of all, as mentioned above, we can still orient $\mho \to V_i$ and $\mho \to V_j$. Then, inspired by that $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ change independently, we can identify the direction between V_i and V_j according to Lemma 1, and we apply our proposed FICP.

A6 Details about the Experiments on Synthetic Datasets

More details about the synthetic datasets are explained in this section, including the implementation details in section A6.1, the results analysis of F_1 and SHD in section A6.2, the complete results of precision and recall in section A6.3, the computational time analysis in section A6.4, the hyperparameter study on the number of hidden features h in section A6.5, the statistical significance test for the results in section A6.6, and the evaluation on dense graph in section A6.7.

A6.1 Implementation Details

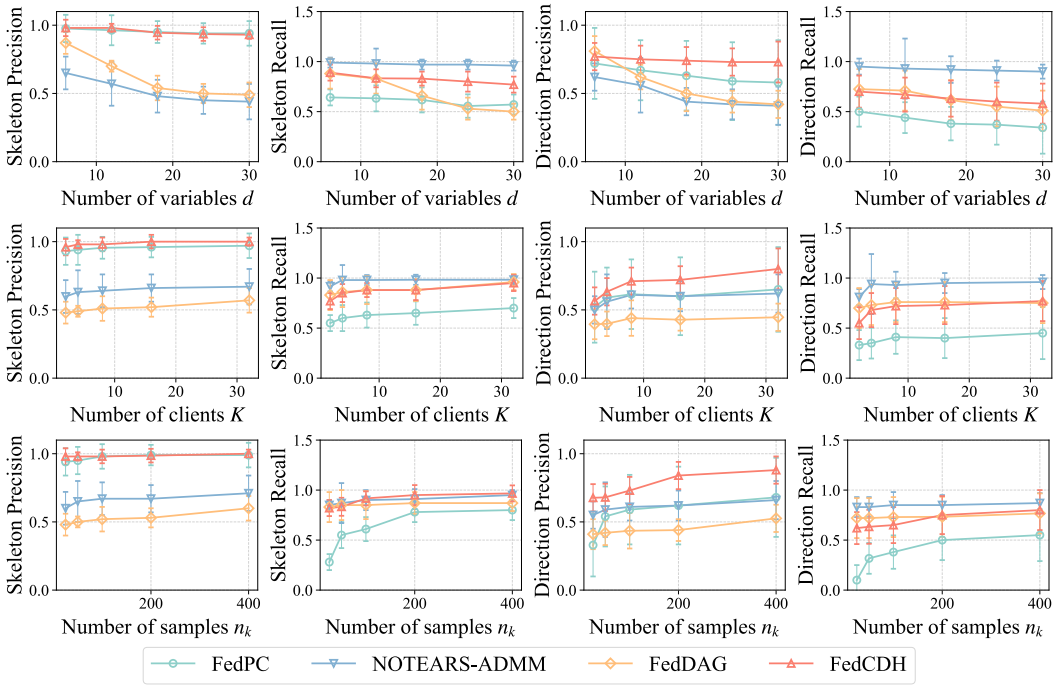
We provide the implementation details of our method and other baseline methods.

- FedDAG (Gao et al., 2022): Codes are available at the author's Github repository `https://github.com/ErdunGAO/FedDAG`. The hyperparameters are set by default.
- NOTEARS-ADMM and NOTEARS-MLP-ADMM (Ng & Zhang, 2022): Codes are available at the author's Github repository `https://github.com/ignavierng/notears-admm`. The hyperparameters are set by default, e.g., we set the threshold level to 0.1 for post-processing.
- FedPC (Huang et al., 2022): Although there is no public implementation provided by the author, considering that it is the only constraint-based method among all the existing works for federated causal discovery, we still compared with it. We reproduced it based on the Causal-learn package `https://github.com/py-why/causal-learn`. Importantly, we follow the paper, set the voting rate as 30% and set the significance level to 0.05.
- FedCDH (Ours): Our method is developed based on the CD-NOD (Huang et al., 2020) and KCI (Zhang et al., 2012) which are publicly available in the Causal-learn package `https://github.com/py-why/causal-learn`. We set the hyperparameter h to 5, and set the significance level for FCIT to 0.05. Our source code has been appended in the Supplementary Materials.

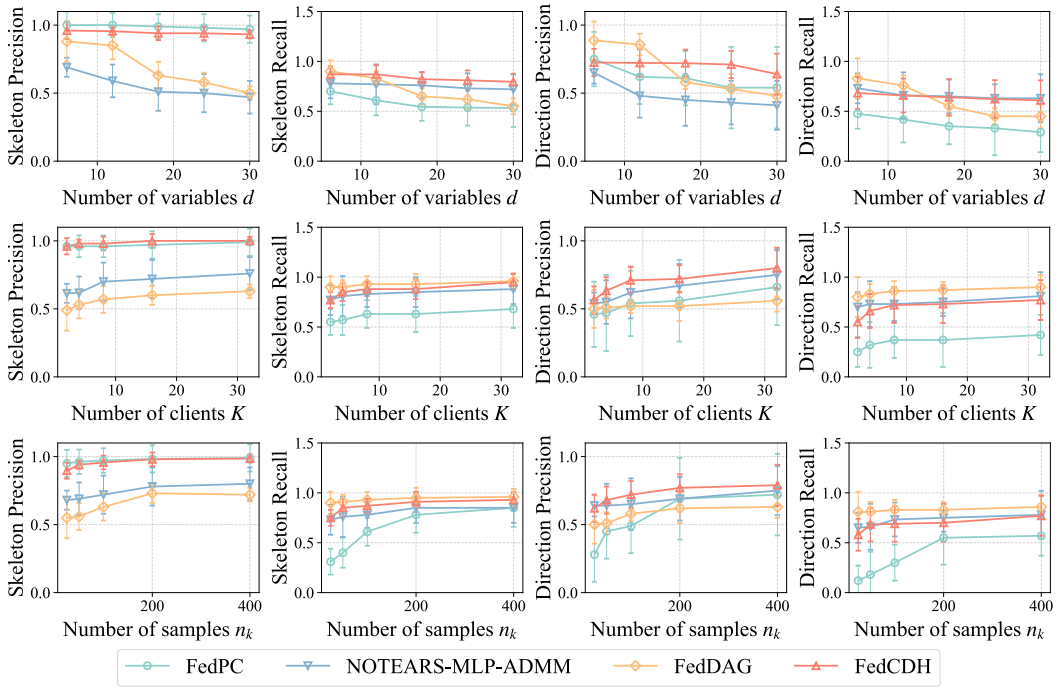
For NOTEARS-ADMM, NOTEARS-MLP-ADMM, and FedDAG, the output is a directed acyclic graph (DAG), while FedPC and our FedCDH may output a completed partially directed acyclic graph (CPDAG). To ease comparisons, we use the simple orientation rules (Dor & Tarsi, 1992) implemented by Causal-DAG (Chandler Squires, 2018) to convert a CPDAG into a DAG. We evaluate both the undirected skeleton and the directed graph, denoted by "Skeleton" and "Direction" as shown in the Figures.

A6.2 Analysis of F_1 and SHD

We have provided the results of F_1 and SHD in the main paper as shown in Figure 3 and Figure 4, here we provide further discussions and analysis.



(a) Precision and recall on linear Gaussian model.



(b) Precision and recall on general functional model.

Figure A2: Results of the synthetic dataset on (a) linear Gaussian model and (b) general functional model. By rows in each subfigure, we evaluate varying number of variables d, varying number of clients K, and varying number of samples n_k. By columns in each subfigure, we evaluate Skeleton Precision (↑), Skeleton Recall (↑), Direction Precision (↑) and Direction Recall (↑).

Table A1: Results of computational time for varying number of variables d, varying number of clients K, and varying number of samples n_k. We report the average and standard deviation over 10 runs. This is the synthetic dataset based on linear Gaussian model.

Data Sizes			**Methods**			
d	K	n_k	FedPC	NOTEARS-ADMM	FedDAG	FedCDH (Ours)
6	10	100	3.87 ± 1.97s	14.10 ± 1.89s	136.92 ± 21.50s	8.14 ± 2.47s
12			32.01 ± 3.54s	28.33 ± 2.46s	321.84 ± 65.94s	62.69 ± 7.77s
18			39.58 ± 4.75s	35.13 ± 2.89s	398.27 ± 149.51s	98.57 ± 9.23s
24			84.05 ± 7.64s	40.01 ± 2.94s	715.80 ± 268.93s	172.11 ± 18.18s
30			94.03 ± 9.48s	56.35 ± 3.91s	1441.13 ± 519.04s	232.35 ± 26.67s
6	2	100	0.72 ± 0.24s	7.04 ± 0.64s	50.38 ± 11.29s	3.88 ± 1.49s
	4		2.07 ± 0.73s	9.07 ± 0.77s	85.08 ± 15.68s	5.24 ± 1.74s
	8		3.64 ± 1.54s	10.80 ± 0.78s	114.81 ± 29.67s	8.01 ± 2.32s
	16		5.79 ± 2.59s	19.40 ± 2.51s	342.34 ± 62.28s	12.60 ± 2.98s
	32		14.08 ± 4.44s	30.56 ± 2.88s	714.06 ± 137.31s	20.30 ± 4.37s
6	10	25	0.48 ± 0.10s	13.06 ± 1.91s	125.77 ± 20.64s	3.75 ± 1.29s
		50	1.47 ± 0.64s	13.75 ± 2.51s	127.25 ± 20.38s	5.74 ± 1.61s
		100	3.87 ± 1.97s	14.10 ± 1.89s	136.92 ± 21.50s	8.14 ± 2.47s
		200	16.52 ± 3.63s	14.68 ± 2.23s	138.67 ± 31.91s	13.78 ± 3.75s
		400	51.10 ± 6.87s	15.90 ± 2.54s	140.37 ± 34.42s	22.86 ± 4.55s

We evaluate variable $d \in \{6, 12, 18, 24, 30\}$ while fixing other variables such as K=10 and n_k=100. We set client $K \in \{2, 4, 8, 16, 32\}$ while fixing others such as d=6 and n_k=100. We let the sample size in one client $n_k \in \{25, 50, 100, 200, 400\}$ while fixing other variables such as d=6 and K=10.

The results of linear Gaussian model are given in Figure 3 and those of general functional model are provided in Figure 4. According to the results, we observe that our FedCDH method generally outperforms all other baselines across different criteria and settings. According to the results of our method on both of the two models, when d increases, the F_1 score decreases and the SHD increases for skeletons and directions, indicating that FCD with more variables might be more challenging. On the contrary, when K and n_k increase, the F_1 score grows and the SHD reduces, suggesting that more joint clients or samples could contribute to better performances for FCD.

In linear Gaussian model, NOTEARS-ADMM and FedPC generally outperform FedDAG. The reason may be that the front two methods were proposed for linear model while the latter one was specially proposed for nonlinear model. *(iv)* In general functional model, FedPC obtained the worst performance compared to other methods in direction F_1 score, possibly due to its strong assumptions on linear model and homogeneous data. FedDAG and NOTEARS-MLP-ADMM revealed poor results regarding SHD, the reasons may be two-fold: they assume nonlinear identifiable model, which may not well handle the general functional model; and both of them are continuous-optimization-based methods, which might suffer from various issues such as convergence and nonconvexity.

A6.3 Results of Precision and Recall

In the main paper, we have only provided the results of F_1 score and SHD, due to the space limit. Here, we provide more results and analysis of the precision and the recall. The results of average and standard deviation are exhibited in Figure A2.

According to the results, we could observe that our FedCDH method generally outperformed all other baseline methods, regarding the precision of both skeleton and direction. However, in the linear Gaussian model, NOTEARS-ADMM generally achieved the best performance regarding the recall although it performed poorly in precision. In the general functional model, when evaluating varying number of clients K and samples n_k, FedDAG performed the best with respect to the recall, however, neither FedDAG nor NOTEARS-MLP-ADMM obtained satisfactory results in the precision.

Table A2: Hyperparameter study on the number of hidden features h. We evaluate the F_1 score, precision, recall, and SHD of both skeleton and direction. We report the average over 10 runs. This is the synthetic dataset based on linear Gaussian model.

h \ Metrics	Skeleton				Direction				Time↓
	F_1 ↑	Precision↑	Recall↑	SHD↓	F_1 ↑	Precision↑	Recall↑	SHD↓	
5	0.916	0.980	0.867	0.9	0.721	0.765	0.683	2.0	8.14s
10	0.916	0.980	0.867	0.9	0.747	0.810	0.700	2.0	8.87s
15	0.907	0.980	0.850	1.0	0.762	0.818	0.717	1.8	10.57s
20	0.889	0.980	0.833	1.2	0.767	0.833	0.717	1.8	12.72s
25	0.896	0.980	0.833	1.1	0.789	0.838	0.750	1.6	20.93s
30	0.896	0.980	0.833	1.1	0.825	0.873	0.783	1.4	37.60s

A6.4 Results of Computational Time

Existing works about federated causal discovery rarely evaluate the computational time when conducting experiments. Actually, it is usually difficult to measure the exact computational time in real life, because of some facts, such as the paralleled computation for clients, the communication time costs between the clients and the server, and so on. However, the computational time is a significant factor to measure the effectiveness of a federated causal discovery method to be utilized in practical scenarios. Therefore, in this section, for making fair comparisons, we evaluate the computational time for each method, assuming that there is no paralleled computation (meaning that we record the computational time at each client and server and then simply add them up) and no extra communication cost (indicating zero time cost for communication).

We evaluate different settings as mentioned above, including varying number of variables d, varying number of clients K, and varying number of samples n_k. We generate data according to linear Gaussian model. For each setting, we run 10 instances, report the average and the standard deviation of the computational time. The results are exhibited in Table A3.

According to the results, we could observe that among the four FCD methods, FedDAG is the least efficient method with the largest time cost. Meanwhile, FedPC, NOTEARS-ADMM and our FedCDH are comparable. In the setting of varying variables, our method exhibited unsatisfactory performance among the three methods. However, in the case of varying variables, NOTEARS-ADMM is the most ineffective method, and in the scenario of varying samples, FedPC is the slowest one among the three methods.

A6.5 Hyperparameter Study

We conduct experiments on the hyperparameter, such as the number of mapping functions or hidden features h. Regarding the experiments in the main paper, we set h to 5 by default. Here in this section, we set $h \in \{5, 10, 15, 20, 25, 30\}$, $d = 6$, $K = 10$, $n_k = 100$ and evaluate the performances. We generate data according to linear Gaussian model. We use the F_1 score, the precision, the recall and the SHD for both skeleton and direction. We also report the runtime. We run 10 instances and report the average values. The experimental results are given in Table A2.

According to the results, we could observe that with the number of hidden features h increasing, the performance of the direction is obviously getting better, while the performance of the skeleton may fluctuate a little bit. Moreover, the computational time is also increasing. When h is smaller than 20, the runtime increases steadily. When h is greater than 20, the runtime goes up rapidly. Importantly, we could see that even when h is small, such as $h = 5$, the general performance of our method is still robust and competitive.

A6.6 Statistical Significance Test

In order to show the statistical significance of our method compared with other baseline methods on the synthetic linear Gaussian model, we report the p values via Wilcoxon signed-rank test (Woolson, 2007). For each baseline method, we evaluate four criteria: Skeleton F1 (S-F1), Skeleton SHD (S-SHD), Direction F1 (D-F1), and Direction SHD (D-SHD).

Table A3: Test result of statistical significance of our FedCDH method compared with other baseline methods. We report the p values via Wilcoxon signed-rank test (Woolson, 2007). This is the synthetic dataset based on linear Gaussian model.

Parameters			[FedCDH vs. FedPC]				[FedCDH vs. NOTEARS-ADMM]				[FedCDH vs. FedDAG]			
d	k	n	S-F_1	S-SHD	D-F_1	D-SHD	S-F_1	S-SHD	D-F_1	D-SHD	S-F_1	S-SHD	D-F_1	D-SHD
6	10	100	0.00	0.05	0.01	0.12	0.00	0.01	0.11	0.10	0.00	0.01	0.01	0.01
12	10	100	0.00	0.01	0.01	0.01	0.00	0.00	0.15	0.00	0.00	0.00	0.11	0.00
18	10	100	0.00	0.01	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.02	0.00
24	10	100	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00
30	10	100	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
6	*2*	100	0.00	0.00	0.01	0.01	0.01	0.00	0.21	0.01	0.00	0.00	0.03	0.00
6	*4*	100	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.00
6	*8*	100	0.00	0.00	0.01	0.02	0.02	0.01	0.03	0.02	0.00	0.00	0.09	0.00
6	*16*	100	0.00	0.01	0.01	0.02	0.00	0.00	0.10	0.03	0.00	0.00	0.07	0.00
6	*32*	100	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.00	0.03	0.00
6	10	*25*	0.00	0.00	0.01	0.01	0.01	0.01	0.26	0.02	0.00	0.00	0.03	0.00
6	10	*50*	0.00	0.01	0.01	0.00	0.01	0.00	0.99	0.03	0.00	0.00	0.02	0.00
6	10	*200*	0.00	0.01	0.01	0.02	0.00	0.00	0.03	0.02	0.00	0.00	0.11	0.01
6	10	*400*	0.00	0.01	0.01	0.01	0.01	0.00	0.03	0.01	0.00	0.01	0.01	0.00

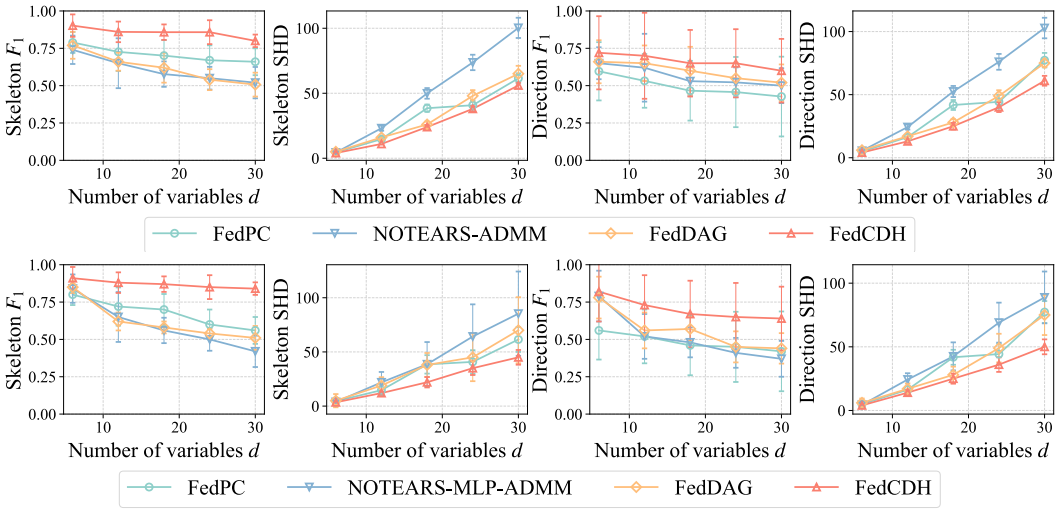


Figure A3: We evaluate on synthetic linear Gaussian model (Top Row) and general functional model (Bottom Row) when the number of edges are two times the number of variables. By columns, we evaluate Skeleton F_1 (↑), Skeleton SHD (↓), Direction F_1 (↑) and Direction SHD (↓).

We set the significance level to 0.05. Those p values higher then 0.05 are underlined. From the results, we can see that the improvements of our method are statistically significant at 5% significance level in general.

A6.7 Evaluation on Dense Graph

As shown in Figure 3 in the main paper, the true DAGs are simulated using the Erdös–Rényi model (Erdős et al., 1960) with the number of edges equal to the number of variables. Here we consider a more dense graph with the number of edges are two times the number of variables.

we evaluate on synthetic linear Gaussian model and general functional model, and record the F_1 score and SHD for both skeleton and directed graphs. All other settings are following the previous ones by default.

According to the results as shown in Figure A3, we can see that our methods still outperformed other baselines in varying number of variables. Interestingly, when the generated graph is more dense, the performance of FedPC will obviously go down for various number of variables.

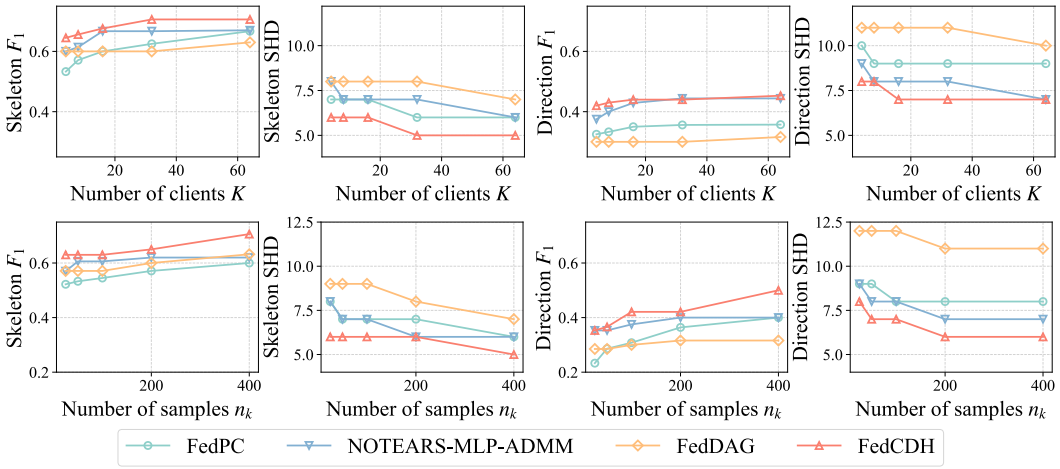


Figure A4: Results of real-world dataset fMRI Hippocampus (Poldrack et al., 2015). By rows, we evaluate varying number of clients K and varying number of samples n_k. By columns, we evaluate Skeleton F_1 (↑), Skeleton SHD (↓), Direction F_1 (↑) and Direction SHD (↓).

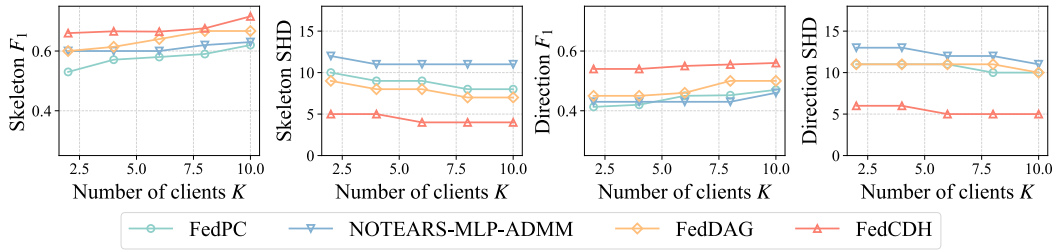


Figure A5: Results of real-world dataset HK Stock Market (Huang et al., 2020). We evaluate varying number of clients K, and we evaluate Skeleton F_1 (↑), Skeleton SHD (↓), Direction F_1 (↑) and Direction SHD (↓).

A7 Details about the Experiments on Real-world Dataset

A7.1 Details about fMRI Hippocampus Dataset

We evaluate our method and the baselines on fMRI Hippocampus (Poldrack et al., 2015). The directions of anatomical ground truth are: PHC → ERC, PRC → ERC, ERC → DG, DG → CA1, CA1 → Sub, Sub → ERC and ERC → CA1. Generally, we follow a similar setting as the experiments on synthetic datasets. For each of them, we use the structural Hamming distance (SHD), the F_1 score as evaluation criteria. We measure both the undirected skeleton and the directed graph. Here, we consider varying number of clients K and varying number of samples in each client n_k.

The results of F_1 score and SHD is given in Figure A4. According to the results, we could observe that our FedCDH method generally outperformed all other baseline methods, across all the criteria listed.

A7.2 Details about HK Stock Market Dataset

We also evaluate on HK stock market dataset (Huang et al., 2020) (See Page 41 for more details about the dataset). The HK stock dataset contains 10 major stocks, which are daily closing prices from 10/09/2006 to 08/09/2010. The 10 stocks are Cheung Kong Holdings (1), Wharf (Holdings) Limited (2), HSBC Holdings plc (3), Hong Kong Electric Holdings Limited (4), Hang Seng Bank Ltd (5), Henderson Land Development Co. Limited (6), Sun Hung Kai Properties Limited (7), Swire Group (8), Cathay Pacific Airways Ltd (9), and Bank of China Hong Kong (Holdings) Ltd (10). Among these stocks, 3, 5, and 10 belong to Hang Seng Finance Sub-index (HSF), 1, 8, and 9

belong to Hang Seng Commerce and Industry Sub-index (HSC), 2, 6, and 7 belong to Hang Seng Properties Sub-index (HSP), and 4 belongs to Hang Seng Utilities Sub-index (HSU).

Here one day can be also seen as one domain. We set the number of clients to be $K \in \{2, 4, 6, 8, 10\}$ while randomly select $n_k = 100$ samples for each client. All other settings are following previous ones by default. The results are provided in Figure A5. According to the results, we can infer that our FedCDH method also outperformed the other baseline methods, across the different criteria.