Reframed GES with a Neural Conditional Dependence Measure

Abstract

In a nonparametric setting, the causal structure is often identifiable only up to Markov equivalence, and for the purpose of causal inference it is useful to learn a graphical representation of the Markov equivalence class (MEC). In this paper, we revisit the Greedy Equivalence Search (GES) algorithm, which is widely cited as a score-based algorithm for learning the MEC of the underlying causal structure. We observe that in order to make the GES algorithm consistent in a nonparametric setting, it is not necessary to design a scoring metric that evaluates graphs. Instead, it suffices to plug in a consistent estimator of a measure of conditional dependence to guide the search. We therefore present a reframing of the GES algorithm, which is more flexible than the standard score-based version and readily lends itself to the nonparametric setting with a general measure of conditional dependence. In addition, we propose a neural conditional dependence (NCD) measure, which utilizes the expressive power of deep neural networks to characterize conditional independence in a nonparametric manner. We establish the optimality of the reframed GES algorithm under standard assumptions and the consistency of using our NCD estimator to decide conditional independence. Together these results justify the proposed approach. Experimental results demonstrate the effectiveness of our method in causal discovery, as well as the advantages of using our NCD measure over kernel-based measures.

1 INTRODUCTION

Causal structure learning is a fundamental problem in various disciplines of science, and flexible solutions to this problem have potentially wide applications [Pearl et al., 2000, Koller and Friedman, 2009, Peters et al., 2017]. In many scenarios, it is expensive or even impossible to perform interventions or randomized experiments in order to reveal the causal relationships. This limitation inspires the need to infer or at least systematically produce plausible hypotheses of causal structures from purely observational data, which is often known as causal discovery. General assumptions relating the data distribution to the unknown causal structure have been leveraged to make causal discovery feasible, including the well-known Markov condition and faithfulness assumption [Spirtes et al., 2000].

Suppose the unknown causal structure can be properly represented by a directed acyclic graph (DAG) over the observed variables. The last one and a half decades have seen a host of nice results on the identifiability of the causal DAG from the observational data distribution, under various parametric or semi-parametric assumptions [Shimizu et al., 2006, Hoyer et al., 2009, Zhang and Hyvarinen, 2009, Bühlmann et al., 2014, Peters and Bühlmann, 2014]. However, in a nonparametric setting, from the observational distribution the causal structure is known to be identifiable only up to Markov equivalence. Despite this limitation, it remains a worthy task to learn a graphical representation of the Markov equivalence class (MEC), known as a completed partial directed acylic graph (CPDAG), for a CPDAG usually reveals some valuable causal information and can be used to guide experimental studies.

Existing methods for causal discovery targeting the CPDAG are roughly categorized into constraint-based and score-based methods. The former uses statistical tests to find conditional (in)dependence relationships in the data and use them as constraints to recover the CPDAG that satisfies them. The PC algorithm [Spirtes et al., 2000] is a well-known exemplar of this approach. The latter formulates the task as an optimization problem by assigning a score to each candidate graph and searching for the one with the optimal score. Regarding the search and optimization strategy, many algorithms solve a combinatorial optimization problem by

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

performing greedy search; on the other hand, starting from Zheng et al. [2018], much recent work tackles the problem through a continuous program [Yu et al., 2019, Lachapelle et al., 2020]. While continuous optimization has advantages in scalability, the global convergence is hard to guarantee by using gradient-based algorithms without implausible assumptions such as strong convexity, especially when the model involves neural networks. In contrast, some search algorithms can be shown to achieve global optimality in the large sample limit even with a relatively sparse search space. One of the best known procedures of this kind is Greedy Equivalence Search (GES) [Chickering, 2002].

In terms of scoring, most score functions are born out of restrictive parametric assumptions on the data distribution. Classical examples include the BIC [Schwarz, 1978] and the BDeu [Geiger and Heckerman, 1994] score. However, when the parametric model is misspecified, which is very common in real data, the optimality of such a score is not guaranteed to reflect the ground truth even in the large sample limit.

In this paper, we explore a simple strategy to produce a nonparametric GES. We observe that in order to make GES consistent in a nonparametric setting, it is not necessary to design a scoring metric that evaluates graphs as a whole. Instead, it suffices to define a certain criterion to guide the search at each step of the procedure. The way we consider in this work is to plug in a consistent estimator of a measure of conditional dependence to provide such guidance. The result is a reframed GES algorithm that is more flexible than the standard score-based version and readily lends itself to the nonparametric setting with a general measure of conditional dependence. This avoids potential model misspecification that commonly occurs in score-based methods. On the other hand, although the reframed GES becomes essentially constraint-based, it retains desirable features of the search strategy of GES and performs significantly better in our experiments than paradigmatic constraint-based methods such as PC.

In addition, we propose a measure of conditional dependence based on a characterization of conditional independence from Daudin [1980] and a novel neural conditional dependence (NCD) estimator which utilizes the expressive power of deep neural networks. Many existing nonparametric measures of conditional dependence are based on kernel methods and leverage characterizations in a Reproducing Kernel Hilbert Space (RKHS), e.g., Gretton et al. [2005]. However, kernel methods suffer from high computational complexity, preventing them from efficient applications in large scale problems. In contrast, our neural network based approach can benefit from a large sample size without a severe compromise in computational time.

We highlight our main contributions as follows:

- We present a reframing of the GES algorithm that can flexibly incorporate a consistent estimator of a general conditional dependence measure.
- We propose a neural conditional dependence (NCD) measure, which utilizes the expressive power of deep neural networks.
- We provide theoretical guarantees on the correctness of the reframed GES algorithm and the consistency of the NCD estimator to measure conditional dependence under mild conditions, and demonstrate the empirical advantages of the resulting method in causal discovery.

2 BACKGROUND AND RELATED WORK

2.1 PRELIMINARIES AND NOTATIONS

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed acyclic graph (DAG) consisting of nodes $\mathbf{V} = (X_1, \ldots, X_d)$, each of which is a (possibly multi-dimensional) random variable, and directed edges $\mathbf{E}$ that connect pairs of nodes. Let $\mathbf{Pa}_i^{\mathcal{G}}$ be the set of parents of node X_i. We denote the joint distribution of $\mathbf{V}$ by $P_{\mathbf{V}}$. A basic problem of causal discovery aims at inferring the unknown causal DAG $\mathcal{G}$ from an independent and identically distributed (i.i.d.) sample from $P_{\mathbf{V}}$. In general, we need assumptions relating the DAG $\mathcal{G}$ and the distribution $P_{\mathbf{V}}$ to make the task possible. A principle adopted by all causal discovery methods is the *causal Markov condition*: $P_{\mathbf{V}}$ is Markovian with respect to DAG $\mathcal{G}$, in the sense that every conditional independence statement entailed by $\mathcal{G}$ according to the standard Markov property of DAGs is true of $P_{\mathbf{V}}$. We also assume the commonly adopted *faithfulness assumption*: $P_{\mathbf{V}}$ is faithful with respect to DAG $\mathcal{G}$, in the sense that every conditional independence statement true of $P_{\mathbf{V}}$ is entailed by $\mathcal{G}$. If two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ entail the same set of conditional independence statements, they are said to be *Markov equivalent*. The set of all DAGs that are Markov equivalent to a DAG $\mathcal{G}$ is called the *Markov equivalence class* (MEC) of $\mathcal{G}$, which can be represented by a completed partially directed acyclic graph (CPDAG). For random variables X, Y and Z, we write $X \perp\!\!\!\perp Y \mid Z$ to mean that X and Y are conditionally independent given Z.

2.2 RELATED WORK ON CAUSAL DISCOVERY

The point that GES can be recast in the spirit of a constraint-based method has been indicated in the literature, most recently by Chickering [2020], but to our knowledge has never been explicitly explored, probably because it appears pointless in parametric settings where well-behaved scores are available. However, as we aim to demonstrate in this paper, the utility of such a perspective on GES is significant in nonparametric settings.

In our experiments we use a number of state-of-the-art causal discovery algorithms for comparison, in addition

to the aforementioned PC and standard GES. One of them is the CAM algorithm [Bühlmann et al., 2014], which decouples the search for the causal ordering from the selection of parents for each variable, by leveraging an additive modelling assumption. Huang et al. [2018] propose a generalized score function (GSF) and apply it in the GES algorithm. Specifically, they transform the statistical decision about conditional independence to a model selection problem for a regression task in a RKHS, define a score based on the penalized log-likelihood for the kernel regression, and then use the score to guide local moves in GES. This work is closely related to ours in that both works are motivated by the goal to develop a nonparametric score to guide the local moves of GES. However, our approach differs from GSF in at least two notable aspects. First, by highlighting the sufficiency of designing a score that enables consistent statistical decisions about conditional independence, we propose a simpler and more flexible way to dispense with parametric assumptions in GES. Second, the specific score we propose is based on neural networks rather than kernels and hence enjoys better computational efficiency when scaling to a large sample size. Another earlier work sharing the similar spirit is the kernel generalized variance (KGV) [Bach and Jordan, 2002], which is also compared in our experiments.

Other methods follow the lead of Zheng et al. [2018], who reformulate the original combinatorial problem into a continuous optimization problem, named NOTEARS, which is solved using the augmented Lagrangian algorithm. Several follow-up works extend NOTEARS to nonlinear causal models, including DAG-GNN [Yu et al., 2019], GraN-DAG [Lachapelle et al., 2020], and Ng et al. [2019], all of which utilize neural networks to model the nonlinear causal relations. In addition, Zhu et al. [2020] adopt policy gradient to search for a DAG with the optimal score.

2.3 RELATED WORK ON CONDITIONAL INDEPENDENCE

Conditional independence plays an important role in many statistics and machine learning problems, ranging from graphical models [Koller and Friedman, 2009] to invariance learning [Arjovsky et al., 2019]. A number of studies were devoted to characterizing conditional independence or developing conditional independence tests. Gretton et al. [2005] introduce the Hilbert-Schmidt independence criterion (HSIC), which is extended by Fukumizu et al. [2007] to cover conditional independence and used for a conditional independence test. Recently, Azadkia and Chatterjee [2021] propose a surprisingly simple nonparametric measure of conditional dependence based on ranking statistics, which we refer to as the Rank Conditional Dependence (RCD) measure and will use later to illustrate the flexibility of our approach. Other works focus on constructing tests of conditional independence by proposing various test statistics, including the kernel conditional independence test [Zhang et al., 2011] and the test based on a generalized covariance measure [Shah and Peters, 2020], among others.

3 REFRAMING THE GES ALGORITHM

3.1 STANDARD GES

The GES algorithm [Chickering, 2002] searches over the space of MECs of DAGs, which are represented by CPDAGs. The connectivity in the search space is given by the independence-map (IMAP) relation: graph $\mathcal{G}$ is an IMAP of graph $\mathcal{G}'$ if every conditional independence entailed by $\mathcal{G}$ is entailed by $\mathcal{G}'$. The standard GES uses a scoring function that assigns a score to every DAG given data, and uses the score of a representative DAG in a MEC as that for the MEC. The search strategy consists of two phases, a phase of forward equivalence search (FES) followed by a phase of backward equivalence search (BES). In FES, the procedure starts with the empty CPDAG (the one with no edges), and moves at each step to a best-scoring CPDAG with one more adjacency (that is an IMAP of the previous CPDAG), until the score cannot be improved by adding more adjacencies. In BES, the procedure starts with the output from FES, and moves at each step to a best-scoring CPDAG with one fewer adjacency (of which the previous CPDAG is an IMAP), until the score cannot be improved by deleting more adjacencies.

We enter some details of FES to highlight the observation that motivates the subsequent reframing. The case of BES is analogous. In FES, each step considers possible insert-one-edge operations on the current CPDAG. Following Chickering [2002], for non-adjacent nodes X_i and X_j in a CPDAG $\mathcal{P}$, and for any subset $\mathbf{T}$ of the neighbors of X_j (i.e., nodes that are connected to X_j by undirected edges) that are not adjacent to X_i, the $Insert(X_i, X_j, \mathbf{T})$ operator modifies $\mathcal{P}$ to obtain $\mathcal{P}'$ by inserting the directed edge $X_i \to X_j$, and for each $T \in \mathbf{T}$, directing the previously undirected edge between T and X_j as $T \to X_j$. If the validity condition in Chickering [2002, Theorem 15] is met, then there is a representative DAG $\mathcal{G}$ in the MEC represented by $\mathcal{P}$ and a representative DAG $\mathcal{G}'$ in the MEC represented by $\mathcal{P}'$, such that $\mathcal{G}'$ is the result of inserting $X_i \to X_j$ in $\mathcal{G}$ (which implies that $\mathcal{P}'$ is an IMAP of $\mathcal{P}$).

In Chickering's (2002) proof of the asymptotic correctness of GES under the causal Markov and faithfulness assumptions, the crucial condition is that the "local" decision between $\mathcal{G}$ and $\mathcal{G}'$ mentioned above asymptotically tracks whether a certain conditional independence relation holds. We make this notion precise in the following definition, in which $\mathbf{D}$ denotes an i.i.d. sample with size n from the joint distribution $P_{\mathbf{V}}$ of $\mathbf{V}$.

Definition 1 (Independence-tracking decision criterion)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be two DAGs over $\mathbf{V}$ that are exactly the same*

except that $\mathcal{G}'$ contains an edge $X_i \to X_j$ that does not appear in $\mathcal{G}$. A decision criterion (based on data $\mathbf{D}$) to choose between $\mathcal{G}$ and $\mathcal{G}'$ (among other options) is independence-tracking if the following two properties hold in the large sample limit:

(i) If $X_j \perp\!\!\!\perp X_i \mid \mathbf{Pa}_j^{\mathcal{G}}$ (according to $P_{\mathbf{V}}$), then the decision criterion favors $\mathcal{G}$ over $\mathcal{G}'$.

(ii) Otherwise, the decision criterion favors $\mathcal{G}'$ over $\mathcal{G}$.

In the standard GES algorithm, a scoring function for DAGs is used to make such local decisions. The induced decision criterion is independence-tracking if the scoring function satisfies the so-called local consistency [Chickering, 2002, Definition 6]. Indeed, Definition 1 is a straightforward generalization of the notion of local consistency for scoring functions. The generalization serves to highlight a simple but important observation: the crucial condition for the optimality of GES can be implemented by a (locally consistent) score function for DAGs, but does not necessitate such a function.

3.2 REFRAMED GES

We now describe a simple alternative way to implement an independence-tracking decision criterion for GES, by using any consistent measure of conditional dependence, in the following sense:

Definition 2 (τ-consistency)**.** *Consider a set of statistics $\mathcal{T} = \{T_n(X, Y|\mathbf{Z}) \mid X, Y \in \mathbf{V}, \mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}\}$ (intended to measure conditional dependence) depending on $\mathbf{D}$. $\mathcal{T}$ is said to be τ-consistent with parameter $\tau > 0$ if for every $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$, the following two conditions hold in the large sample limit:*

(i) If $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ (according to $P_{\mathbf{V}}$), then $T_n(X, Y|\mathbf{Z}) < \tau$.

(ii) Otherwise, $T_n(X, Y|\mathbf{Z}) > \tau$.

For our purpose, the following sufficient condition for the τ-consistency in Definition 2 is useful. All proofs are deferred to Appendix B.

Proposition 1. *Suppose for every $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$, $T_*(X, Y|\mathbf{Z}) \geq 0$ is a quantity depending on $P_{\mathbf{V}}$ such that*

$$T_*(X, Y|\mathbf{Z}) = 0 \text{ if and only if } X \perp\!\!\!\perp Y \mid \mathbf{Z}.$$

Let $\hat{T}_n(X, Y|\mathbf{Z})$ form a set of statistics indexed by $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$. If $\hat{T}_n(X, Y|\mathbf{Z}) \to T_*(X, Y|Z)$ in probability as $n \to \infty$ for every $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$, then there exists $\tau > 0$ such that $\{\hat{T}_n(X, Y|\mathbf{Z})\}$ is τ-consistent.

This proposition provides a way to construct a τ-consistent set of statistics. One first defines a population quantity that takes the boundary value if and only if the conditional independence in question holds, and then constructs a consistent estimator for this quantity given an i.i.d. sample. The aforementioned measures including HSIC [Fukumizu et al., 2007] and RCD [Azadkia and Chatterjee, 2021] were both developed along this line. In the next section, we will propose a new measure of conditional dependence based on a neural network implementation.

It is worth noting the essential difference between our defined τ-consistent statistics and conditional independence tests. We note that the condition in Proposition 1 indicates that when $X \perp\!\!\!\perp Y \mid Z$, the statistic $\hat{T}_n(X, Y|\mathbf{Z})$ converges to 0 in probability, i.e., $\hat{T}_n = o_p(1)$. By contrast, in a typical conditional independence test, one usually uses a test statistic that, under the null hypothesis of conditional independence, follows an asymptotic null distribution, which is then used to develop a decision rule. This means that when $X \perp\!\!\!\perp Y \mid Z$, the test statistic is stochastically bounded, i.e., $O_p(1)$, but not necessarily $o_p(1)$. Therefore, it is in general non-trivial to define a τ-consistent statistic from a conditional independence test.

With a τ-consistent $\hat{T}(X, Y|\mathbf{Z})$, it is straightforward to implement an independence-tracking decision criterion to be used in GES. Specifically, at each step in FES, to each (valid) operator $Insert(X_i, X_j, \mathbf{T})$ we assign $\hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}})$ as its "local score" (where $\mathcal{G}$ is the DAG representing the current CPDAG induced by the operator), and apply the operator with the highest local score (indicating conditional dependence), unless all remaining valid insert operators yield a local score lower than the threshold τ. Similarly, at each step in BES, to each (valid) operator $Delete(X_i, X_j, \mathbf{H})$, we assign $\hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}})$ as its local score (where $\mathcal{G}$ is the DAG representing the CPDAG the operator would produce), and apply the operator with the lowest local score (indicating conditional independence), unless all remaining valid delete operators yield a score greater than τ. The update step of FES is summarized in Algorithm 1, and the dual update step of BES is given in Appendix A.

We call the GES algorithm with these update steps the *reframed GES*. Unlike the standard GES, this reframed GES does not optimize a global score for MECs. However, by using a suitable local score for choosing operators to apply (or to stop), the local decision criterion remains independence-tracking, and as a result the asymptotic optimality of the reframed GES algorithm is still guaranteed, as stated in the following theorem.

Theorem 2. *Under the causal Markov and faithfulness assumptions, the reframed GES procedure using a τ-consistent $\hat{T}(X, Y|\mathbf{Z})$ recovers the MEC of the true graph in the large sample limit.*

Algorithm 1 The update step in the reframed FES

Input: the current CPDAG $\mathcal{P}$, sample $\mathbf{D}$, a list of valid insert operators **INS**, statistics $\hat{T}(X, Y|\mathbf{Z})$, threshold τ
Output: the next CPDAG $\mathcal{P}'$

```
Set s = 0 and I = NULL.
for Insert(X_i, X_j, T) ∈ INS do
    Let G be the representative DAG of P corresponding to
    Insert(X_i, X_j, T).
    Evaluate Score(X_i, X_j, T) = T̂(X_i, X_j | Pa_j^G).
    if Score(X_i, X_j, T) > s then
        Let s = Score(X_i, X_j, T) and I =
        Insert(X_i, X_j, T).
    end if
end for
if s > τ then
    Apply operator I to obtain P'.
else
    Keep P' = P (and terminate FES).
end if
return P'
```

4 NEURAL CONDITIONAL DEPENDENCE MEASURE

In this section, we propose a novel measure of conditional dependence. Let X, Y, and Z be three random variables taking values in $\mathbb{R}^{d_X}$, $\mathbb{R}^{d_Y}$, and $\mathbb{R}^{d_Z}$, respectively, and we assume that their joint distribution is absolutely continuous with respect to Lebesgue measure with density p_* defined on $\mathbb{R}^{d_X+d_Y+d_Z}$. The conditional independence between X and Y given Z is defined by $p_*(x, y, z) = p_*(x|z)p_*(y|z)$ for all x, y, z with $p_*(z) > 0$ [Dawid, 1979].

The following lemma from Daudin [1980] characterizes the conditional independence, which has given rise to several hypothesis testing methods. Let L^2_Z, L^2_{XZ}, and L^2_{YZ} be the spaces of square integrable functions of Z, (X, Z), and (Y, Z), repsectively, e.g., $L^2_{XZ} = \{f : \mathbb{R}^{d_X+d_Z} \to \mathbb{R} \mid \mathbb{E}[f(X, Z)^2] < \infty\}$.

Lemma 3 (Daudin [1980])**.** *The random variables X and Y are conditionally independent given Z if and only if*

$$\mathbb{E}[f(X, Z)g(Y, Z)] = 0,$$

for all $f \in L^2_{XZ}$ and $g \in L^2_{YZ}$ such that $\mathbb{E}[f(X, Z)|Z] = 0$ and $\mathbb{E}[g(X, Z)|Z] = 0$.

At the population level, given a ground truth density p_*, we propose the following measure of conditional dependence between X and Y given Z:

$$S(X, Y|Z) = \sup_{f,g} \rho^2(f(X, Z) - h^*(Z), g(Y, Z) - l^*(Z)) \quad (1)$$

where $f \in L^2_{XZ}$ and $g \in L^2_{YZ}$ are test functions, $h^*(Z) = \mathbb{E}[f(X, Z)|Z]$, $l^*(Z) = \mathbb{E}[g(Y, Z)|Z]$, and $\rho(X_1, X_2) = \mathsf{cov}(X_1, X_2)/\sqrt{\mathsf{var}(X_1)\mathsf{var}(X_2)}$ denotes the Pearson correlation coefficient of two random variables X_1 and X_2. The reason of using the correlation coefficient rather than the covariance is that after normalization by the variances, the characteristic is bounded between $[-1, 1]$. This makes the measure well-defined in a bounded range and the computation of its subsequent sample version numerically stable.

Based on Lemma 3, we have the following simple theorem which characterizes the property of the measure (1) and establishes an equivalence condition between the measure and conditional independence.

Theorem 4. *For all p_*, we have $S(X, Y|Z) \in [0, 1]$ and $S(X, Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$.*

Having defined the measure $S(X, Y|Z)$, we now make the computation tractable. We use deep neural network classes to parametrize the test functions f, g and the conditional expectations h, l in (1). Formally, we write f_θ, g_ϕ, h_ω, l_ψ, where the subscripts denote the parameters of the corresponding neural networks. We then exploit the approximation

$$\sup_{\theta,\phi} \rho^2(f_\theta(X, Z) - h_{\omega^*}(Z), g_\phi(Y, Z) - l_{\psi^*}(Z)), \quad (2)$$

where $h_{\omega^*}(z) = h^*(z)$ and $l_{\psi^*}(z) = l^*(z)$. According to the universal approximation theorem of neural networks [Hornik et al., 1989], equation (2) can approximate the true measure (1) with arbitrary accuracy by choosing the appropriate network architecture. Since here we mainly focus on the statistical property of the estimator proposed below, we ignore the small approximation error (i.e., the gap between (2) and (1)) in the analysis for simplicity.

Next, we present a consistent estimator of $S(X, Y|Z)$. Let $\mathbf{D} = \{(x_i, y_i, z_i), i = 1, \ldots, n\}$ be the collection of i.i.d. copies of $(X, Y, Z) \sim p_*$. Our estimator of $S(X, Y|Z)$ is given by

$$\hat{S}_n(X, Y|Z) = \sup_{\theta,\phi} \hat{\rho}^2\big(f_\theta(X, Z) - h_{\hat{\omega}}(Z), g_\phi(Y, Z) - l_{\hat{\psi}}(Z)\big), \quad (3)$$

where $\hat{\rho}$ is the sample correlation coefficient based on data $\mathbf{D}$, and

$$\begin{aligned} \hat{\omega} &= \operatorname*{argmin}_{\omega} \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i, z_i) - h_\omega(z_i))^2, \\ \hat{\psi} &= \operatorname*{argmin}_{\psi} \frac{1}{n} \sum_{i=1}^{n} (g_\phi(y_i, z_i) - l_\psi(z_i))^2, \end{aligned} \quad (4)$$

are the estimators of ω^* and ψ^*.

Remark. The estimators in (4) based on regression come from the fact that $\mathbb{E}[f(X, Z)|Z] = \operatorname{argmin}_h \mathbb{E}[f(X, Z) - h(Z)]^2$, which is proved in Appendix B.

We call the proposed estimator (3) the *neural conditional dependence (NCD)* estimator and its population version (2) the NCD measure. Since (3) solves a bilevel optimization problem involving regression problems (4), we adopt an

Algorithm 2 Computing the NCD score

Input: sample $\mathbf{D}$, horizon T_t, T_r, initial $\theta, \phi, \omega, \psi$
Output: NCD score

for $t_t = 1, 2, \ldots, T_t$ **do**
  for $t_r = 1, 2, \ldots, T_r$ **do**
    Update ω by descending $\sum_i \nabla_\omega (f_\theta(x_i, z_i) - h_\omega(z_i))^2$
    Update ψ by descending $\sum_i \nabla_\psi (g_\phi(y_i, z_i) - l_\psi(z_i))^2$
  end for
  Update θ, ϕ by ascending the gradient of
    $\hat{\rho}^2(f_\theta(X, Z) - h_\omega(Z), g_\phi(Y, Z) - l_\psi(Z))$
end for
Compute $\hat{s} = \hat{\rho}^2(f(X, Z) - h(Z), g(Y, Z) - l(Z))$
return $\hat{s}$

alternating gradient descent scheme to obtain the NCD estimator. The procedure is summarized in Algorithm 2, where we alternately update the test functions f_θ, g_ϕ and nonlinear regressors h_ω, l_ψ for T_t and T_r steps, respectively.

To study the asymptotic behavior of the proposed estimator, we assume the following regularity conditions, all of which are mild assumptions commonly adopted in the literature.

C1. The parameter spaces $\theta \in \Theta$, $\phi \in \Phi$, $\omega \in \Omega$, and $\psi \in \Psi$ are compact.

C2. $f_\theta(x, z)$, $g_\phi(y, z)$, $h_\omega(z)$, and $l_\psi(z)$ are continuous with respect to the corresponding parameters and data x, y, z.

C3. $f_\theta(x, z)$, $g_\phi(y, z)$, $h_\omega(z)$, and $l_\psi(z)$ are dominated square integrable, i.e., there exists a dominating function $F(x, z)$ such that $|f_\theta(x, z)| \leq F(x, z)$ for all θ and $\mathbb{E}[F(X, Z)]^2 < \infty$.

C4. For all θ, ϕ, there exist unique $\omega^*(\theta) \in \Omega$ and $\psi^*(\phi) \in \Psi$ such that $h_{\omega^*}(z) = h^*(z)$ and $l_{\psi^*}(z) = l^*(z)$ almost surely, respectively.

The following theorem establishes the consistency of $\hat{S}_n(X, Y|Z)$ as an estimator of the population measure $S(X, Y|Z)$.

Theorem 5. *Under the regularity conditions C1-C4, as $n \to \infty$, we have $\hat{S}_n(X, Y|Z) \to S(X, Y|Z)$ in probability.*

Finally, we apply the NCD measure to causal discovery through the reframing of GES in the previous section. We plug the estimator $\hat{S}_n$ into the reframed GES procedure as the "local score". Based on Theorems 4 and 5, by applying the sufficient condition in Proposition 1, we know that $\hat{S}_n(X, Y|\mathbf{Z})$ satisfies the τ-consistency in Definition 2 with some $\tau > 0$. Then Theorem 2 implies the asymptotic correctness of our method to recover the true MEC.

In addition, to demonstrate the flexibility of our reframed GES algorithm in incorporating various conditional dependence measures, we will also test a version of our procedure using the RCD measure recently proposed by Azadkia and Chatterjee [2021], because it is very easy to compute. The RCD estimator can be shown to satisfy the τ-consistency and hence suits our framework well. For completeness, we provide a description of RCD in Appendix C.

5 EXPERIMENTS

In this section, we compare our proposed method with various existing state-of-the-art causal discovery approaches on both synthetic and pseudo-real data sets. Baseline methods include score-based methods using BIC [Chickering, 2002], KGV [Bach and Jordan, 2002], and GSF [Huang et al., 2018]; a constraint-based method, PC algorithm [Spirtes et al., 2000]; a method based on structural causal model, CAM [Bühlmann et al., 2014]; as well as the emerging methods in the continuous optimization paradigm including NOTEARS [Zheng et al., 2018], DAG-GNN [Yu et al., 2019], and GraN-DAG [Lachapelle et al., 2020]. The details of the experimental settings and hyperparameters (including the choice of τ) of baseline methods and ours are given in Appendix D.[1]

The causal discovery performance is evaluated using three metrics: the structural hamming distance (SHD), the structural interventional distance (SID) [Peters and Bühlmann, 2015] and the F1 score. Since our method and many baseline approaches return a CPDAG representing an MEC, both SHD and SID are evaluated between the learned and ground-truth CPDAGs. Then the SHD is the smallest number of edge additions, deletions, and reversals to convert the estimated CPDAG into the true CPDAG. The SID counts the number of pairs (X_i, X_j) such that the interventional distribution $p(x_j|do(X_i = x))$ would be miscalculated if we chose the parent adjustment set from the estimated graph. We report the SIDs corresponding to the best and worst DAG in the learned MEC. The F1 score is defined as the harmonic mean of the precision and the recall. In general, a lower SHD or SID and a higher F1 score indicate a better estimate.

5.1 SYNTHETIC DATA

As mentioned in previous sections, when a data set does not satisfy the additive Gaussian noise assumption, many existing methods such as BIC, CAM, NOTEARS, and GraN-DAG suffer from model misspecification and thus may lead to misleading results. In contrast, nonparametric methods like GSF and ours in principle will not be affected. Here we consider the well-known post nonlinear (PNL) causal models [Zhang and Hyvarinen, 2009]. A general PNL model expresses each variable X_i as

$$X_i = g_{i,2}(g_{i,1}(\mathbf{Pa}_i) + N_i),\ i = 1, \ldots, d,$$

[1]The code will be released in the final version of the paper.

Setting	GP (1k)		GP (5k)		MULT (1k)		MULT (5k)	
Methods	SHD	F1 score	SHD	F1 score	SHD	F1 score	SHD	F1 score
NCD	**5.6±2.5**	**0.63±0.14**	**4.2±2.3**	**0.71±0.14**	6.2±2.9	0.59±0.08	5.6±2.4	0.60±0.08
RCD	9.0±0.7	0.41±0.07	8.4±1.1	0.53±0.08	7.4±2.1	0.51±0.09	**3.2±1.3**	**0.67±0.07**
PC	8.8±1.6	0.36±0.15	7.2±2.4	0.50±0.16	7.6±1.7	0.44±0.15	4.6±1.8	0.57±0.13
BIC	7.0±2.8	0.49±0.20	6.0±2.5	0.59±0.17	4.2±2.9	0.65±0.09	4.4±3.4	0.62±0.11
KGV	8.5±1.1	0.37±0.08	7.5±0.5	0.51±0.06	9.0±1.9	0.35±0.14	7.2±0.7	0.47±0.07
CAM	6.0±3.5	0.50±0.26	7.2±3.7	0.52±0.22	10.8±1.8	0.09±0.07	11.2±2.3	0.13±0.15
NOTEARS	11.4±0.9	0.06±0.08	11.6±0.9	0.06±0.08	24.8±3.8	0.36±0.07	23.6±4.7	0.37±0.07
DAG-GNN	11.0±1.7	0.00±0.00	11.4±1.8	0.03±0.07	16.4±2.6	0.37±0.13	13.6±3.4	0.40±0.10
GraN-DAG	10.6±1.1	0.05±0.06	12.2±1.8	0.12±0.04	8.6±2.6	0.54±0.12	10.2±1.9	0.51±0.08
GSF	6.4±3.5	0.55±0.19	>12h		**3.0±1.1**	**0.67±0.06**	>12h	

Table 1: SHD and F1 score on PNL data sets with 10 nodes, 2 expected degrees, and 1000 and 5000 samples.

Setting	GP (1k)		GP (5k)		MULT (1k)		MULT (5k)	
Methods	SHD	F1 score	SHD	F1 score	SHD	F1 score	SHD	F1 score
NCD	**28.4±3.6**	**0.55±0.05**	**24.6±3.8**	**0.58±0.08**	**29.2±4.6**	**0.52±0.07**	29.8±5.1	**0.57±0.09**
RCD	32.8±2.2	0.39±0.11	32.6±4.7	0.44±0.10	31.4±4.5	0.43±0.14	**27.2±3.3**	0.54±0.06
PC	37.6±1.3	0.18±0.09	36.0±2.7	0.26±0.05	36.2±1.8	0.23±0.06	34.4±1.5	0.32±0.08
BIC	33.0±2.1	0.45±0.06	30.8±3.3	0.50±0.09	30.8±5.6	0.43±0.09	35.0±3.9	0.39±0.07
KGV	37.8±0.7	0.20±0.08	34.2±3.9	0.33±0.07	37.2±1.6	0.27±0.02	37.4±2.2	0.31±0.06
CAM	33.0±5.6	0.42±0.13	30.6±3.4	0.50±0.11	35.2±2.8	0.25±0.07	34.4±6.5	0.31±0.15
NOTEARS	38.8±1.9	0.13±0.05	38.4±1.8	0.13±0.05	39.0±1.6	0.33±0.04	39.0±1.9	0.34±0.07
DAG-GNN	39.2±1.3	0.03±0.02	39.2±2.3	0.05±0.09	37.8±2.4	0.26±0.10	39.6±1.1	0.25±0.12
GraN-DAG	34.0±7.9	0.18±0.09	35.4±6.9	0.30±0.13	37.4±3.2	0.20±0.08	37.0±3.5	0.27±0.09
GSF	34.0±3.0	0.39±0.05	>12h		31.6±3.2	0.38±0.09	>12h	

Table 2: SHD and F1 score on PNL data sets with 10 nodes, 8 expected degrees, and 1000 and 5000 samples.

where $\mathbf{Pa}_i$ contains the direct causes of X_i, N_i is the exogenous noise variable, and $g_{i,1}$ and $g_{i,2}$ are nonlinear transformations.

To synthesize a data set, we first randomly generate a ground-truth DAG $\mathcal{G}$ following the Erdős-Rényi (ER) graph model and then generate data following $\mathcal{G}$ and two types of PNL models that were also considered in Lachapelle et al. [2020]. The first one, called *PNL-GP*, samples $g_{i,1}$ independently from a Gaussian process with bandwidth one, takes $g_{i,2}$ as the sigmoid function, and $N_i \sim Laplace(0, b_i)$ with $b_i \sim \mathcal{U}[0, 1]$. All root variables in PNL-GP are sampled from $\mathcal{U}[-1, 1]$. The second one, named *PNL-MULT*, takes $g_{i,1}(x) = \log(sum(x))$ where $sum(x)$ takes the sum of all components of a vector x, $g_{i,2}(\cdot) = \exp(\cdot)$, and $N_i \sim |\mathcal{N}(0, \sigma_i^2)|$ with $\sigma_i^2 \sim \mathcal{U}[0, 1]$. All root variables in PNL-MULT are sampled from $\mathcal{U}[0, 2]$. This model is adapted from Zhang et al. [2015].

Tables 1 and 2 present the results of SHD and F1 score on sparse and dense graphs with 10 nodes respectively, where the error bars represent the standard deviations across 5 data sets per setting. The results of SID are basically consistent, which are deferred to Appendix E due to the space limit. Additional results on graphs with 20 nodes are also presented in Appendix E. We see that in general, the reframed GES algorithm with our own NCD or the adopted RCD (shown in the first two lines of all tables) performs the best across all settings, except in the sparse PNL-MULT data where GSF is the best. The advantages of our methods on the more challenging dense graphs are more significant than those on sparse ones. In most cases, NCD outperforms RCD, though RCD produces excellent results on PNL-MULT with a larger sample. From the perspective of implementation, RCD may be favored over NCD in terms of fewer hyperparameters and less computational cost. In addition, our methods exhibit similarly good performances across different ground-truth models, while most other methods tend to perform well on at most one setting, which indicates the robustness of our nonparametric approach against different distributions.

GSF, as another kernel-based nonparametric score, performs very well on PNL-MULT with a sparse structure, but is less competitive in other settings. Note that we only report the results of GSF using 1000 samples, because even for the sparse graph, it takes around 17 hours for a single run with

Method	SHD	SID	F1 score
NCD	**2.6±2.9**	**[2.0±3.7, 14.2±2.8]**	**0.73±0.15**
PC	3.0±1.3	[7.4±4.3, 14.9±3.7]	0.57±0.11
KGV	4.1±1.6	[3.7±3.3, 17.8±2.8]	0.58±0.11
DAG-GNN	4.3±2.3	[4.1±3.9, 15.6±4.1]	0.59±0.27
GSF	4.7±3.0	[4.3±4.2, 15.6±3.0]	0.61±0.17
BIC	4.7±0.9	[4.8±3.8, 16.2±2.7]	0.57±0.00

Table 3: Results on 10 multi-dimensional data sets.

Method	SynTReN	CHILD
NCD	**30.0±5.8**	**16.8±3.3**
RCD	**30.9±4.8**	**14.0±5.8**
PC	37.4±4.1	22.6±9.4
BIC	65.8±10.8	32.8±16.8
NOTEARS	99.8±14.4	23.6±1.9
DAG-GNN	38.5±5.1	28.8±3.2
GraN-DAG	58.7±10.0	17.2±2.1
KGV	39.9±8.0	>12h
GSF	52.1±8.9	>12h

Table 4: SHD on pseudo-real data.

5000 samples compared to around 18 minutes with 1000 samples. In contrast, our NCD-based method can benefit from a larger sample size while taking similar computational time as with a smaller sample (both within 3 minutes in the sparse case). KGV leads to inferior performance in all settings. The standard GES with the linear-Gaussian BIC score sometimes performs well on PNL-MULT; a possible reason is that when the noise variance $\sigma_i^2 \sim \mathcal{U}[0,1]$ happens to be small, the PNL-MULT model behaves similarly to a linear-Gaussian model, leading to a case with minor misspecification for BIC. This may also partly account for the fact that PC performs better on PNL-MULT than on PNL-GP; that is, a PNL-MULT data set can be similar to linear-Gaussian data which would satisfy the model assumption made in the hypothesis testing. CAM performs better on PNL-GP than on PNL-MULT and achieves the best SID in one case, as shown in Appendix E. The continuous optimization methods are inferior on these PNL data sets, which could be explained by their misspecification of the model.

In addition, we evaluate our methods in a multi-dimensional scenario where each node may have more than one dimension. Note that our NCD estimator can be readily applied to the multi-dimensional setup by adjusting the input dimension of the test functions, while the rank-based RCD measure unfortunately cannot be directly applied here. Some of the baseline methods, including CAM, NOTEARS, and GraN-DAG are not applicable to the multi-dimensional case, so we do not compare with them in this setting. We use 10 synthetic data sets from Huang et al. [2018], each with 5 nodes and a sample size of 1000. As shown in Table 3, our approach outperforms the baseline methods in all three metrics. PC, KGV, and GSF are the second best performing methods in terms of SHD, SID, and F1 score, respectively, though they all give inferior performance in other metrics.

5.2 PSEUDO-REAL DATA

Although the synthetic data sets from PNL models can expose the model misspecification problem in many existing methods, they differ from the additive noise setup only by the nonlinearity $g_{i,2}$, and hence amount to relatively mild cases of misspecification. In this section, we consider a pseudo-real data set sampled from the SynTReN generator [Van den Bulcke et al., 2006] where there is no guarantee at all for model specification. We evaluate on the 10 data sets sampled by Lachapelle et al. [2020], each with 20 nodes and a small sample size of 500. In addition, we consider a real Bayesian network, CHILD network (with 20 nodes), and randomly generate 3000 samples following the PNL-GP model introduced in the previous section.

As shown in Table 4, on SynTReN, most baseline methods perform poorly, indicating a potentially severe violation of their model assumptions. Our reframed GES with NCD and with RCD obtain the best SHDs. The results on SynTReN suggest the potential advantage of nonparametric causal discovery methods in real applications where model misspecification is common and possibly grave. Moreover, our methods also obtain the best performance on the real graph structure CHILD. Note that on this large data set with 3000 samples, the kernel-based methods GSF and KGV face serious computational challenges in that they take more than 12 hours for a single run. Therefore, we expect our method to exhibit even more advantages than the kernel-based methods in large-scale scenarios.

6 CONCLUSION

In this work, we presented a reframed GES algorithm that works with a measure of conditional dependence rather than a scoring metric for graphs. This way the algorithm is easily applicable in a nonparametric setting with a theoretical guarantee. We also proposed a neural conditional dependence (NCD) measure based on a deep neural network implementation, and established its theoretical properties that make it suitable for the reframed GES. The resulting causal discovery algorithm was shown in our experiments to be superior or competitive in comparison to a number of state-of-the-art methods. It also enjoys a significant advantage over kernel-based nonparametric methods in large-scale settings, since the latter are usually infeasible when the sample size is relatively large. For future work, we plan to explore a continuous optimization formulation of causal discovery based on such nonparametric conditional dependence measures.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 2021.

Francis R Bach and Michael I Jordan. Learning graphical models with mercer kernels. In *Advances in neural information processing systems*, volume 15, pages 1009–1016, 2002.

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

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

Max Chickering. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. PMLR, 2020.

JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, volume 20, pages 489–496, 2007.

Dan Geiger and David Heckerman. Learning gaussian networks. In *Uncertainty Proceedings 1994*, pages 235–243. Elsevier, 1994.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL `https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf`.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Robert I Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *J. Mach. Learn. Res.*, 21:37–1, 2020.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B1QRgziT-`.

Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

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

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

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 06 2020. doi: 10.1214/19-AOS1857. URL `https://doi.org/10.1214/19-AOS1857`.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1): 1–12, 2006.

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

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

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in neural information processing systems*, 2018.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1g2skStPB`.

A REFRAMED BES

We present the dual update step of the reframed BES in Algorithm 3.

Algorithm 3 The update step in the reframed BES

Input: the current CPDAG $\mathcal{P}$, sample $\mathbf{D}$, a list of valid delete operators $\mathbf{DEL}$, statistics $\hat{T}(X, Y|\mathbf{Z})$, threshold τ

Output: the next CPDAG $\mathcal{P}'$

```
Set s = 0 and I = NULL.
for Delete(X_i, X_j, H) ∈ DEL do
    Let G be the DAG induced by the operator Delete(X_i, X_j, H) that is a representative of the CPDAG the
    operator would produce.
    Evaluate Score(X_i, X_j, H) = T̂(X_i, X_j | Pa_j^G).
    if Score(X_i, X_j, H) < s then
        Let s = Score(X_i, X_j, H) and I = Delete(X_i, X_j, H).
    end if
end for
if s < τ then
    Apply operator I to obtain P'.
else
    Keep P' = P (and terminate BES).
end if
return P'
```

B PROOFS

B.1 PROOF OF PROPOSITION 1

Define the following two sets of tuples

$$\mathcal{A} = \{(X, Y, \mathbf{Z}) : X, Y \in \mathbf{V}, Z \subseteq \mathbf{V} \text{ such that } X \perp\!\!\!\perp Y \mid \mathbf{Z}\};$$

$$\mathcal{B} = \{(X, Y, \mathbf{Z}) : X, Y \in \mathbf{V}, Z \subseteq \mathbf{V} \text{ such that } X \not\perp\!\!\!\perp Y \mid \mathbf{Z}\}.$$

We know from the proposition condition that for every $(X, Y, \mathbf{Z}) \in \mathcal{A}$, $T_*(X, Y|\mathbf{Z}) = 0$ and for every $(X, Y, \mathbf{Z}) \in \mathcal{B}$, $T_*(X, Y|\mathbf{Z}) > 0$. For the number of nodes is finite, the cardinalities of both sets are finite. Then we know $m_0 = \min_{(X,Y,\mathbf{Z})\in\mathcal{B}} T_*(X, Y|\mathbf{Z}) > 0$. Let τ be any number in $(0, m_0)$.

In case (1), by the consistency of $\hat{T}_n$ to T_*, we have $\mathbb{P}(\hat{T}_n(X, Y|\mathbf{Z}) > \tau) \to 0$ as $n \to \infty$.

In case (2), we have as $n \to \infty$, $\hat{T}_n(X, Y|\mathbf{Z}) \xrightarrow{p} T_*(X, Y|\mathbf{Z}) \leq m_0$, where $\xrightarrow{p}$ stands for converging in probability, which means for all $\epsilon > 0$, $\mathbb{P}(|\hat{T}_n(X, Y|\mathbf{Z}) - T_*(X, Y|\mathbf{Z})| < \epsilon) \to 1$. By the arbitrariness of ϵ, let $\epsilon < T_*(X, Y|\mathbf{Z}) - \tau$. Then we have $\{|\hat{T}_n - T_*| < \epsilon\} \subseteq \{T_* - \epsilon < \hat{T}_n\} \subseteq \{\tau < \hat{T}_n\}$. This implies $\mathbb{P}(|\hat{T}_n - T_*| < \epsilon) \leq \mathbb{P}(\hat{T}_n > \tau)$. Therefore, we have $\mathbb{P}(\hat{T}_n > \tau) \to 1$ as $n \to \infty$, which concludes the proof.

B.2 PROOF OF THEOREM 2

The proof is essentially the same as the proof for the asymptotic correctness of the standard GES with a locally consistent scoring function [Chickering, 2002], except that the role played by the local consistency of the scoring function is now played by the τ-consistency of $\hat{T}$. We first show that in the large sample limit, the output of the reframed FES is a CPDAG $\mathcal{P}$ that satisfies the Markov condition with the true distribution $P_{\mathbf{V}}$. Suppose for the sake of contradiction that $P_{\mathbf{V}}$ is not Markov to $\mathcal{P}$, which means that $P_{\mathbf{V}}$ is not Markov to any DAG $\mathcal{G}$ in (the MEC represented by) $\mathcal{P}$. It follows that there exists a pair of distinct variables X_i, X_j such that they are not adjacent in $\mathcal{G}$ and X_i is a non-descendant of X_j in $\mathcal{G}$, but X_i and X_j are not independent given $\mathbf{Pa}_j^{\mathcal{G}}$ according to $P_{\mathbf{V}}$. However, since $\hat{T}$ is τ-consistent, in the large sample limit

$\hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}}) > \tau$, which means that the reframed FES would not have stopped with $\mathcal{P}$ but would have moved to another CPDAG with an added adjacency between X_i and X_j. Contradiction.

Next we show that if the reframed BES starts with a CPDAG that is Markov to $P_{\mathbf{V}}$, then in the large sample limit it will output the CPDAG that is both Markov and faithful to $P_{\mathbf{V}}$, which represents the true MEC by the causal Markov and faithfulness assumptions. Suppose for the sake of contradiction that the reframed BES ends with a CPDAG $\mathcal{P}$ that is not faithful to $P_{\mathbf{V}}$. Note that $\mathcal{P}$ would still be Markov to $P_{\mathbf{V}}$. If not, since the reframed BES starts with a CPDAG that is Markov to $P_{\mathbf{V}}$, there must have been a step where it moved from a CPDAG that is Markov to $P_{\mathbf{V}}$ to one that is not. Denote the latter by $\mathcal{P}'$. It follows that the local score for the operator $Delete(X_i, X_j, \mathbf{H})$ leading to $\mathcal{P}'$ — which is equal to $\hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}'})$, for some $\mathcal{G}'$ in (the MEC represented by) $\mathcal{P}'$ — is smaller than τ (in the large sample limit) even though X_i and X_j are not independent given $\mathbf{Pa}_j^{\mathcal{G}'}$ according to $P_{\mathbf{V}}$. This contradicts the τ-consistency of $\hat{T}$.

Thus the reframed BES ends with a $\mathcal{P}$ that is Markov but not faithful to $P_{\mathbf{V}}$. Let $\mathcal{H}$ denote the true CPDAG, which by assumption is both Markov and faithful to $P_{\mathbf{V}}$. Then $\mathcal{P}$ is an IMAP of $\mathcal{H}$. By Theorem 4 in Chickering [2002], there is a $\mathcal{P}'$ with one more adjacency than $\mathcal{P}$ has such that $\mathcal{P}$ is also an IMAP of $\mathcal{P}'$. It follows that there is a DAG $\mathcal{G}'$ representing $\mathcal{P}'$ and a $\mathcal{G}$ representing $\mathcal{P}$ such that $\mathcal{G}'$ and $\mathcal{G}$ are the same except for an edge $X_i \to X_j$ in $\mathcal{G}'$ but not in $\mathcal{G}$, and $X_i \perp\!\!\!\perp X_j \mid \mathbf{Pa}_j^{\mathcal{G}'}$ according to $P_{\mathbf{V}}$. Since $\hat{T}$ is τ-consistent, we have $\hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}'}) < \tau$ in the large sample limit. But this means that the reframed BES would not have stopped at $\mathcal{P}$ but would have continued to some other CPDAG. A contradiction.

Therefore, the reframed FES followed by the reframed BES will output the true CPDAG $\mathcal{H}$ in the large sample limit.

B.3 PROOF OF THEOREM 4

It is obvious that a correlation coefficient always lies in $[-1, 1]$, so $S(X, Y|Z) \in [0, 1]$. For the second half of the theorem, note that

$$\rho(f(X,Z) - h^*(Z), g(Y,Z) - l^*(Z)) = \frac{\mathbb{E}[(f(X,Z) - h^*(Z))(g(Y,Z) - l^*(Z))]}{\sqrt{\mathbb{E}[f(X,Z) - h^*(Z)]^2 \mathbb{E}[g(Y,Z) - l^*(Z)]^2}}.$$

We have

$$\{f \in L^2_{XZ} : \mathbb{E}[f(X,Z)|Z] = 0\} = \{\tilde{f} | \tilde{f}(X,Z) = f(X,Z) - \mathbb{E}[f(X,Z)|Z], f \in L^2_{XZ}\} := \mathcal{E}_{XZ},$$

$$\{g \in L^2_{YZ} : \mathbb{E}[g(Y,Z)|Z] = 0\} = \{\tilde{g} | \tilde{g}(Y,Z) = g(Y,Z) - \mathbb{E}[g(Y,Z)|Z], g \in L^2_{YZ}\} := \mathcal{E}_{YZ}.$$

We thus have $S(X, Y|Z) = 0$ if and only if

$$\mathbb{E}[\tilde{f}(X,Z)\tilde{g}(Y,Z)] = 0 \quad \forall \tilde{f} \in \mathcal{E}_{XZ}, \tilde{g} \in \mathcal{E}_{YZ}.$$

Then by Lemma 3, we have $S(X, Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$.

B.4 PROOF OF THEOREM 5

Rewrite the NCD estimator as

$$\hat{S}_n = \sup_{\theta \in \Theta, \phi \in \Phi} \frac{\hat{\mathbb{E}}^2[(f_\theta(X,Z) - h_{\hat{\omega}}(Z)) \cdot (g_\phi(Y,Z) - l_{\hat{\psi}}(Z))]}{\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\hat{\omega}}(Z)]^2 \cdot \hat{\mathbb{E}}[g_\phi(Y,Z) - l_{\hat{\psi}}(Z)]^2},$$

where $\hat{\mathbb{E}}$ denotes the sample mean given the sample $\mathbf{D} = \{(x_i, y_i, z_i), i = 1, \ldots, n\}$, e.g.,

$$\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\hat{\omega}}(Z)]^2 = \frac{1}{n}\sum_{i=1}^{n}[f_\theta(x_i, z_i) - h_{\hat{\omega}}(z_i)]^2.$$

By the continuous mapping theorem, it suffices to show the following three convergence statements uniformly over $\theta \in \Theta$ and $\phi \in \Phi$:

(i) $\sup_{\theta \in \Theta, \phi \in \Phi} \left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\hat{\omega}}(Z)) \cdot (g_\phi(Y,Z) - l_{\hat{\psi}}(Z))] - \mathbb{E}[(f_\theta(X,Z) - h_{\omega^*}(Z)) \cdot (g_\phi(Y,Z) - l_{\psi^*}(Z))]\right| \xrightarrow{p} 0$;

(ii) $\sup_{\theta \in \Theta} \left|\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\hat{\omega}}(Z)]^2 - \mathbb{E}[f_\theta(X,Z) - h_{\omega^*}(Z)]^2\right| \xrightarrow{p} 0$;

(iii) $\sup_{\phi\in\Phi}\left|\hat{\mathbb{E}}[g_\phi(Y,Z)-l_{\hat{\psi}}(Z)]^2-\mathbb{E}[g_\phi(Y,Z)-l_{\psi^*}(Z)]^2\right|\xrightarrow{p}0.$

Proof of (ii) and (iii). By the triangular inequality, we have

$$\begin{aligned}&\sup_{\theta\in\Theta}\left|\hat{\mathbb{E}}[f_\theta(X,Z)-h_{\hat{\omega}}(Z)]^2-\mathbb{E}[f_\theta(X,Z)-h_{\omega^*}(Z)]^2\right|\\ \le&\sup_{\theta\in\Theta}\left|\hat{\mathbb{E}}[f_\theta(X,Z)-h_{\hat{\omega}}(Z)]^2-\hat{\mathbb{E}}[f_\theta(X,Z)-h_{\omega^*}(Z)]^2\right|+\sup_{\theta\in\Theta}\left|\hat{\mathbb{E}}[f_\theta(X,Z)-h_{\omega^*}(Z)]^2-\mathbb{E}[f_\theta(X,Z)-h_{\omega^*}(Z)]^2\right|\end{aligned}\tag{5}$$

where the second term on the right-hand side vanishes in probability as $n\to\infty$ by applying the uniform law of large numbers [Jennrich, 1969, Theorem 2].

We then write the first term as follows:

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n\left([f_\theta(x_i,z_i)-h_{\hat{\omega}}(z_i)]^2-[f_\theta(x_i,z_i)-h_{\omega^*}(z_i)]^2\right)\right|\\ =&\left|\frac{1}{n}\sum_{i=1}^n[h_{\hat{\omega}}(z_i)-h_{\omega^*}(z_i)]\,[h_{\hat{\omega}}(z_i)+h_{\omega^*}(z_i)-2f_\theta(x_i,z_i)]\right|\\ \le&\left|\frac{1}{n}\sum_{i=1}^n 2f_\theta(x_i,z_i)\,[h_{\hat{\omega}}(z_i)-h_{\omega^*}(z_i)]\right|+\left|\frac{1}{n}\sum_{i=1}^n\left[h_{\hat{\omega}}^2(z_i)-h_{\omega^*}^2(z_i)\right]\right|.\end{aligned}\tag{6}$$

We recall the definitions

$$\begin{aligned}\omega^*(\theta)&=\operatorname*{argmin}_{\omega\in\Omega}\mathbb{E}[f_\theta(X,Z)-h_\omega(Z)]^2\\ \hat{\omega}(\theta)&=\operatorname*{argmin}_{\omega\in\Omega}\frac{1}{n}\sum_{i=1}^n[f_\theta(x_i,z_i)-h_\omega(z_i)]^2.\end{aligned}$$

By the uniform law of large numbers, for all $\theta\in\Theta$, we have as $n\to\infty$ that

$$\sup_{\omega\in\Omega}\left|\hat{\mathbb{E}}[f_\theta(X,Z)-h_\omega(Z)]^2-\mathbb{E}[f_\theta(X,Z)-h_\omega(Z)]^2\right|\xrightarrow{p}0.$$

Further by condition *C4*, we have for all $\theta\in\Theta$, as $n\to\infty$, $\hat{\omega}(\theta)\xrightarrow{p}\omega^*(\theta)$. Let K be an arbitrary compact subset of $\mathbb{R}^{d_Z}$. Because of the compactness of Θ and the Lipschitz continuity of $h_{\hat{\omega}(\theta)}(z)$ and $h_{\omega^*(\theta)}(z)$ over $(\theta,z)\in\Theta\times K$, we have

$$\sup_{\theta\in\Theta,z\in K}|h_{\hat{\omega}(\theta)}(z)-h_{\omega^*(\theta)}(z)|\xrightarrow{p}0$$

as $n\to\infty$, where $\|\cdot\|$ stands for the Euclidean norm. By the continuous mapping theorem, we have as $n\to\infty$,

$$\sup_{\theta\in\Theta,z\in K}|h_{\hat{\omega}(\theta)}^2(z)-h_{\omega^*(\theta)}^2(z)|\xrightarrow{p}0.\tag{7}$$

Next, we show the second term in (6) vanishes in probability. Given an arbitrary $r>0$, let $B_r=\{z\in\mathbb{R}^{d_Z}:\|z\|\le r\}$. Let $B_r^c=\mathbb{R}^{d_Z}\setminus B_r$ be its complement. We have

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\left[h_{\hat{\omega}}^2(z_i)-h_{\omega^*}^2(z_i)\right]\right|&\le\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left|h_{\hat{\omega}}^2(z_i)-h_{\omega^*}^2(z_i)\right|\\ &=\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left[\left|h_{\hat{\omega}}^2(z_i)-h_{\omega^*}^2(z_i)\right|\mathbf{1}_{\{z_i\in B_r\}}+\left|h_{\hat{\omega}}^2(z_i)-h_{\omega^*}^2(z_i)\right|\mathbf{1}_{\{z_i\in B_r^c\}}\right]\\ &\le\sup_{\theta\in\Theta,z\in B_r}|h_{\hat{\omega}}^2(z)-h_{\omega^*}^2(z)|+\frac{2}{n}\sum_{i=1}^n H^2(z_i)\mathbf{1}_{\{z_i\in B_r^c\}},\end{aligned}\tag{8}$$

where the second term in the upper bound (8) comes from the dominated integrable condition in *C3* with a dominating function $H(z)$. By taking $n \to \infty$, the first term in (8) vanishes in probability by (7), and the second term in (8) becomes $\mathbb{E}[H^2(Z)\mathbf{1}_{\{Z\in B_r^c\}}]$. By the dominated convergence theorem, further by letting $r \to \infty$, $\mathbb{E}[H^2(Z)\mathbf{1}_{\{Z\in B_r^c\}}] \to 0$. Thus, we have as $n \to \infty$

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left[h_{\hat{\omega}}^2(z_i) - h_{\omega^*}^2(z_i)\right]\right| \xrightarrow{p} 0. \tag{9}$$

Last, we show the first term in (6) vanishes in probability. Again, we consider an arbitrary radius $r > 0$ and a compact ball $B_r' = \{(x,z) \in \mathbb{R}^{d_X+d_Z} : \|(x,z)\| \le r\}$. Note that $f_\theta(x,z)$ is continuous and hence is uniformly bounded for all $\theta \in \theta$ and $(x,z) \in B_r'$. Then

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i,z_i)\left[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)\right]\right| \\
\le&\frac{1}{n}\sum_{i=1}^{n}|f_\theta(x_i,z_i)[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)]|\mathbf{1}_{\{(x_i,z_i)\in B_r'\}} + \frac{1}{n}\sum_{i=1}^{n}|f_\theta(x_i,z_i)[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)]|\mathbf{1}_{\{(x,z)\in B_r'^c\}} \\
\le& M\sup_{\theta\in\Theta, z\in K}|h_{\hat{\omega}(\theta)}(z) - h_{\omega^*(\theta)}(z)| + \frac{2}{n}\sum_{i=1}^{n}F(x,z)H(z)\mathbf{1}_{\{(x,z)\in B_r'^c\}}
\end{aligned}$$

where $|f_\theta(x,z)| \le M$ for all $\theta \in \Theta$ and $(x,z) \in B_r'$, and $F(x,z)$ and $H(z)$ are dominating functions for $f_\theta(x,z)$ and $h_\omega(z)$ respectively. Similar to the arguments above, we have as $n \to \infty$

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i,z_i)\left[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)\right]\right| \xrightarrow{p} 0. \tag{10}$$

Then by combining convergence results (9) and (10) and recalling the upper bounds (5) and (6), we have as $n \to \infty$, (ii) holds. Similarly we can show (iii). □

Proof of (i). By the triangular inequality, we have

$$\begin{aligned}
&\sup_{\theta\in\Theta,\phi\in\Phi}\left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\hat{\omega}}(Z))\cdot(g_\phi(Y,Z) - l_{\hat{\psi}}(Z))] - \mathbb{E}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right| \\
\le&\sup_{\theta\in\Theta,\phi\in\Phi}\left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\hat{\omega}}(Z))\cdot(g_\phi(Y,Z) - l_{\hat{\psi}}(Z))] - \hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right| \\
&+\sup_{\theta\in\Theta,\phi\in\Phi}\left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))] - \mathbb{E}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right|,
\end{aligned} \tag{11}$$

where the second term on the right-hand side vanishes in probability by the uniform law of large numbers. By some calculations we know that the first term of (11) is upper bounded by

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i,z_i)[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)]\right| + \sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n} g_\phi(y_i,z_i)[l_{\hat{\psi}}(z_i) - l_{\psi^*}(z_i)]\right| \\
&+\sup_{\theta\in\Theta,\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\left[h_{\hat{\omega}}(z_i)l_{\hat{\psi}}(z_i) - h_{\omega^*}(z_i)l_{\psi^*}(z_i)\right]\right|,
\end{aligned}$$

where all three terms converge to 0 in probability as $n \to \infty$. Therefore, the left-hand side of (11) vanishes in probability, leading to (i). □

B.5 PROOF OF THE STATEMENT IN REMARK 4

We recall that $h^*(Z) = \mathbb{E}[f(X,Z)|Z]$. The goal is to show that $h^* = \operatorname{argmin}_{h\in L_Z^2}\mathbb{E}[f(X,Z) - h(Z)]^2$ almost surely.

For all $h \in L_Z^2$, we have

$$\begin{aligned}\mathbb{E}[f(X,Z)-h(Z)]^2 &= \mathbb{E}[(f(X,Z)-h^*(Z))+(h^*(Z)-h(Z))]^2 \\ &= \mathbb{E}[f(X,Z)-h^*(Z))]^2+\mathbb{E}[h^*(Z)-h(Z)]^2+\mathbb{E}[(f(X,Z)-h^*(Z))(h^*(Z)-h(Z))].\end{aligned} \tag{12}$$

Note that the cross term in the second line of (12) can be simplified using the law of total expectation as follows

$$\begin{aligned}\mathbb{E}[(f(X,Z)-h^*(Z))(h^*(Z)-h(Z))] &= \mathbb{E}[\mathbb{E}[(f(X,Z)-h^*(Z))(h^*(Z)-h(Z))]|Z] \\ &= \mathbb{E}[(h^*(Z)-h(Z))\mathbb{E}[f(X,Z)-h^*(Z)|Z]] \\ &= \mathbb{E}[(h^*(Z)-h(Z))(\mathbb{E}[f(X,Z)|Z]-h^*(Z))] \\ &= 0.\end{aligned}$$

Then (12) becomes

$$\mathbb{E}[f(X,Z)-h(Z)]^2 = \mathbb{E}[f(X,Z)-h^*(Z))]^2+\mathbb{E}[h^*(Z)-h(Z)]^2 \geq \mathbb{E}[f(X,Z)-h^*(Z))]^2,$$

where the equality holds if and only if $h(Z)=h^*(Z)$ almost surely.

C RANK CONDITIONAL DEPENDENCE MEASURE

In this section, we briefly introduce the RCI and one may refer to Azadkia and Chatterjee [2021] for details. Consider a random variable Y and two random vectors X and Z, following the joint distribution p_*. Let μ be the law of Y. The following quantity measures the degree of conditional dependence of Y and Z given X:

$$T(X,Y|Z) = \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{P}(Y \geq t|X,Z)|Z))d\mu(t)}{\int \mathbb{E}(\mathrm{Var}(1_{\{Y\geq t\}}|Z))d\mu(t)},$$

which satisfies $T \in [0,1]$ and $T=0$ if and only if $X \perp\!\!\!\perp Y \mid Z$, according to Azadkia and Chatterjee [2021, Theorem 2.1].

Now consider an i.i.d. sample $(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)$ from p_*. For each $i=1\ldots,n$, let $N(i)$ be the index j such that Z_j is the nearest neighbor of Z_i with respect to the Euclidean metric on $\mathbb{R}^{d_Z}$. Let $M(i)$ be the index j such that (X_j,Z_j) is the nearest neighbor of (X_i,Z_i) in $\mathbb{R}^{d_X+d_Z}$. Let R_i be the rank of Y_i. The RCI score is

$$\hat{T}_n(X,Y|Z) = \frac{\sum_{i=1}^n(\min(R_i,R_{M(i)})-\min(R_i,R_{N(i)}))}{\sum_{i=1}^n(R_i-\min(R_i,R_{N(i)}))},$$

which is a consistent estimator of $T(X,Y|Z)$, according to Azadkia and Chatterjee [2021, Theorem 2.2].

D EXPERIMENTAL DETAILS

D.1 IMPLEMENTATIONS OF BASELINE METHODS

All baseline methods were run with the publicly available code from the authors' websites as listed below, expect KGV which we implemented by ourselves:

- GES: We adopt the FGES implementation from `https://github.com/eberharf/fges-py`. Note that all the methods using GES as the search procedure, including our proposed NCD, the adopted RCD, as well as the previous BIC and KGV, are based on the same implementation for searching with the only difference being the updating rule at each step. NCD and RCD follow the reframed GES update step in Algorithms 1 and 3; BIC and KGV follow the standard GES update.
- BIC: The linear-Gaussian BIC score is included in the above FGES implementation.
- KGV: It adopted a Gaussian kernel with kernel width equal to twice of median distance between points in input space.
- PC: An implementation is available through the `py-causal` package at `https://github.com/bd2kccd/py-causal`. We choose SEM-BIC test with significance level 0.05 for PC.

- GSF: An implementation is available at the first author's github repository `https://github.com/Biwei-Huang/Generalized-Score-Functions-for-Causal-Discovery`.
- CAM: An implementation is available through the CRAN R package repository at `https://cran.r-project.org/web/packages/CAM`.
- NOTEARS: The code is available at the first author's github repository `https://github.com/xunzheng/notears`.
- DAG-GNN: The code is available at the first author's github repository `https://github.com/fishmoon1234/DAG-GNN`.
- GraN-DAG: The code is available at the first author's github repository `https://github.com/kurowasan/GraN-DAG`.

In the experiments, we mostly used the default hyperparameters found in the authors' codes unless otherwise stated.

D.2 EVALUATION DETAILS

In the evaluation, both SHD and SID are computed using functions corresponding to CPDAGs in the Causal Discovery Toolbox [Kalainathan et al., 2020]. F1 score, which depends on the precision and recall, involves summarizing the number of correctly estimated edges. Directed edges in the ground-truth CPDAG are deemed correctly estimated if the learned CPDAG contains exactly the same directed edge and are deemed incorrectly otherwise. Undirected edges in the ground-truth CPDAG are converted to two directed edges in the adjacency matrix. When the learned CPDAG contains exactly the same undirected edge, both converted directed edges are correctly estimated. One directed edge and no edge in the learned CPDAG are deemed as correctly estimating 1 and 0 edge, respectively.

D.3 EXPERIMENTAL DETAILS AND HYPERPARAMETERS

Since our model is based on deep neural network, it is sensitive to the choice of hyperparameters, which is also observed in other neural network based causal discovery methods such as Lachapelle et al. [2020]. The hyperparameters in our NCD method include the threshold τ to control the sparsity level (number of edges) of the learned structure, the learning rates of the optimization steps in Algorithm 2, and the neural network architectures (i.e., the number of hidden layers and hidden neurons per layer) for the test functions and nonlinear regressors. The principle of tuning τ is that a larger τ leads to a sparser DAG. To tune τ, one needs an initial guess of the true sparsity, e.g., from domain expert knowledge, and tunes down τ if the learned DAG is much sparser than expected and vice versa.

We use multilayer perceptrons (MLPs) to represent the test functions and regressors. A test function MLP has several blocks each of which consists of a fully connected layer and a ReLU activation function; a regressor MLP further adds batch normalization before the ReLU layer in each block. We adopt spectral normalization [Miyato et al., 2018] in all networks to guarantee the Lipschitz continuity of them. The neural network models and optimization are implemented based on Pytorch. We use Adam optimizer with full batch gradients and a learning rate of 0.01 for both test functions and regressors. We take the training steps $T_t = 20$ and $T_r = 5$ for test functions and regressors respectively. Since test functions serve as transformations to detect correlation (which is a simpler task) while regressors need to fit the data (which is a more complex task), we keep the architecture of test functions fixed with 2 layers and 20 neurons per layer, while only tune the network size of regressors for different data. The above listed hyperparameters remain unchanged across all settings. For different ground-truth causal models with varying dimensions and degrees, we only tune the threshold τ and the network depth and width. For the RCD, the only tuning parameter is the threshold τ. We listed the specific hyperparameters for the our experiments in Table 5.

Setting	depth	width	τ_{NCD}	τ_{RCD}
PNL data with degree 2	3	40	0.005	0.05
PNL data with degree 8	3	80	0.0001	0.001
Multi-dimensional data	3	50	0.01	-
SynTReN	4	100	0.3	0.5

Table 5: Hyperparameters

Our NCD computation involves randomness coming from the neural network initialization (and stochastic optimization if adopted). Next, we introduce a metric based on the proposed NCD estimator to select among the random runs and to some extent guide hyperparameter tuning in an unsupervised manner (i.e., without access to the ground-truth structure). Given a candidate DAG $\mathcal{G}$, let $\mathbf{Pa}_i^{\mathcal{G}}$ and $\mathbf{Nd}_i^{\mathcal{G}}$ be the sets of parents and non-descendants of node X_i, respectively. We propose the following global score to characterize how well the observational data satisfies the conditional independence relations entailed by $\mathcal{G}$:

$$S_g(\mathcal{G}) = \frac{1}{d}\sum_{i=1}^{d} \hat{S}_n(X_i, \mathbf{Nd}_i^{\mathcal{G}} | \mathbf{Pa}_i^{\mathcal{G}}).$$

Apparently we have $S_g(\mathcal{G}) \in [0, 1]$. According to Theorems 4 and 5, we have $S_g(\mathcal{G}) \xrightarrow{p} 0$ as $n \to \infty$ if and only if $\mathcal{G}$ satisfies the Markov condition to the data distribution $P_{\mathbf{V}}$. Hence a candidate DAG $\mathcal{G}$ with a smaller global score $S_g(\mathcal{G})$ is regarded as a better estimate in the large sample limit. For each data set in our experiments, we run our reframed GES algorithm with NCD with two different random initializations and select the one with the lower global score.

E ADDITIONAL EXPERIMENTAL RESULTS

We present the results of SID on the PNL data sets in Tables 6-7 as a supplement to Tables 1-2. We can see that the SIDs are mostly consistent with the SHDs and F1 scores. In general, our NCD or RCD are among the best SID methods. In the sparse PNL-MULT data where GSF is the best with a smaller sample; in the dense graph (with degree 8), CAM performs well on both GP and MULT with a larger sample.

In addition, we present the results of all methods on the PNL data sets with 20 nodes, 2 expected degrees and 5000 samples in Tables 8-9. We observe that our reframed GES with NCD or RCD performs among the best methods in this setting.

Method	GP (1k)	GP (5k)	MULT (1k)	MULT (5k)
NCD	**[11.2±7.0, 24.6±11.4]**	**[6.8±4.8, 19.2±9.5]**	[10.6±5.7, 23.6±10.6]	[12.8±5.7, 25.4±6.1]
RCD	[18.6±4.7, 30.2±7.5]	[17.4±2.9, 26.6±6.5]	[14.0±5.6, 29.8±11.3]	**[4.8±1.4, 22.0±7.8]**
PC	[17.0±6.6, 27.2±6.3]	[15.6±7.4, 27.2±7.9]	[18.4±8.3, 32.6±6.9]	[8.0±5.7, 23.2±5.9]
BIC	[15.0±8.8, 24.8±7.9]	[15.2±8.6, 23.4±8.4]	[7.6±7.3, 23.4±5.3]	[10.0±8.1, 25.2±4.7]
KGV	[18.5±4.0, 27.5±6.2]	[15.5±4.4, 29.0±2.2]	[15.3±6.1, 30.3±11.1]	[10.0±3.0, 28.6±10.4]
CAM	[10.8±6.2, 22.6±9.6]	[17.0±8.9, 28.6±9.8]	[27.6±13.6, 43.0±13.9]	[22.0±12.3, 37.2±16.5]
NOTEARS	[21.2±3.4, 26.2±5.0]	[21.0±3.6, 26.6±4.7]	[15.0±4.2, 21.4±15.9]	[12.8±5.9, 17.0±9.1]
DAG-GNN	[23.6±6.9, 27.4±5.8]	[30.4±11.0, 36.0±9.2]	[16.0±5.6, 23.2±9.9]	[16.0±3.9, 30.0±11.8]
GraN-DAG	[27.0±7.5, 38.2±8.4]	[31.4±8.5, 41.8±7.3]	[14.0±5.7, 27.0±7.0]	[12.4±8.6, 25.6±6.5]
GSF	[14.2±8.6, 24.4±8.9]	-	**[5.0±1.4, 21.6±6.7]**	-

Table 6: SID on PNL datasets with 10 nodes, 2 expected degrees, and 1000 and 5000 samples.

Method	GP (1k)	GP (5k)	MULT (1k)	MULT (5k)
NCD	**[56.6±11.5, 67.6±3.9]**	**[58.8±8.2, 66.0±3.7]**	**[59.2±10.0, 68.8±3.7]**	**[51.4±7.6, 69.0±3.9]**
RCD	[75.4±5.8, 75.4±5.8]	[73.6±4.3, 74.4±3.8]	[67.8±14.1, 77.0±4.5]	[53.2±6.6, 73.2±4.8]
PC	[78.2±10.8, 85.0±4.6]	[76.6±7.7, 82.4±3.6]	[72.2±4.5, 80.8±5.8]	[69.4±11.5, 78.4±5.6]
BIC	[69.8±7.1, 73.2±8.9]	[68.0±7.9, 68.8±8.3]	[67.8±6.7, 78.2±3.1]	[69.8±9.5, 77.4±4.7]
KGV	[74.6±7.7, 83.0±5.3]	[77.0±4.4, 79.6±5.0]	[67.2±3.4, 89.6±0.8]	[66.4±9.4, 87.8±3.0]
CAM	[65.6±10.3, 78.6±4.0]	**[54.6±13.4, 75.6±7.8]**	[56.8±4.1, 83.2±3.5]	**[51.8±27.0, 83.0±3.2]**
NOTEARS	[75.8±1.8, 78.4±1.8]	[75.4±1.3, 78.0±1.2]	[63.4±6.9, 83.6±3.5]	[62.0±7.8, 83.6±4.3]
DAG-GNN	[84.8±4.9, 89.6±0.5]	[86.8±2.6, 89.0±1.7]	[63.4±12.7, 83.2±3.0]	[72.4±5.4, 80.0±2.3]
GraN-DAG	[72.6±22.8, 84.6±2.4]	[68.8±15.8, 78.6±5.6]	[67.4±6.6, 85.6±2.9]	[62.4±6.3, 82.6±3.7]
GSF	[69.6±10.3, 77.4±9.0]	-	[66.6±7.8, 81.0±3.8]	-

Table 7: SID on PNL datasets with 10 nodes, 8 expected degrees, and 1000 and 5000 samples.

Setting	SHD	SID	F1
NCD	**9.2±4.1**	**[37.0±29.0,65.8±52.1]**	**0.69±0.10**
RCD	15.0±1.6	[53.5±28.9, 89.2±34.7]	0.45±0.09
PC	13.2±1.9	[42.6±15.0, 123.0±47.8]	0.58±0.07
BIC	11.6±1.8	[53.8±16.1,93.6±22.9]	0.60±0.03
CAM	**7.4±3.8**	**[38.8±21.5, 38.8±21.5]**	**0.75±0.11**
NOTEARS	23.6±2.9	[120.4±30.7, 120.4±30.7]	0.06±0.02
DAG-GNN	21.4±3.4	[105.2±35.5, 105.2±35.5]	0.11±0.00
GraN-DAG	13.4±3.1	[79.8±31.8, 79.8±31.8]	0.50±0.14

Table 8: Results on PNL-GP data with 20 nodes, 2 expected degrees and 5000 samples.

Setting	SHD	SID	F1
NCD	**10.2±1.9**	[30.4±4.2, 72.6±14.9]	0.59±0.05
RCD	**8.2±1.6**	**[22.6±5.1, 69.2±12.3]**	**0.63±0.04**
PC	10.4±1.2	[16.6±6.7,88.4±18.5]	0.59±0.03
BIC	12.0±2.3	[24.8±6.2, 65.8±13.4]	0.58±0.05
CAM	22.0±3.5	[99.8±12.5, 99.8±12.5]	0.20±0.08
NOTEARS	29.0±2.3	[52.8±13.4, 52.8±13.4]	0.41±0.09
DAG-GNN	36.4±13.1	[58.2±14.6, 58.2±14.6]	0.35±0.09
GraN-DAG	18.2±3.86	[75.4±10.1, 75.4±10.1]	0.25±0.10

Table 9: Results on PNL-MULT data with 20 nodes, 2 expected degrees and 5000 samples.