Reframed GES with a Neural Conditional Dependence Measure (Supplementary material)

Xinwei Shen[1] **Shengyu Zhu**[2] **Jiji Zhang**[3] **Shoubo Hu**[2] **Zhitang Chen**[2]

[1]Hong Kong University of Science and Technology
[2]Huawei Noah's Ark Lab
[3]Hong Kong Baptist University

A REFRAMED BES

We present the dual update step of the reframed BES in Algorithm 1.

Algorithm 1 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 $\tau$
Output: the next CPDAG $\mathcal{P}'$
Set $s = 0$ and $I = \texttt{NULL}$.
for $Delete(X_i, X_j, \mathbf{H}) \in \mathbf{DEL}$ do
    Let $\mathcal{G}$ be the DAG induced by the operator $Delete(X_i, X_j, \mathbf{H})$ that is a representative of the CPDAG the operator would produce.
    Evaluate $Score(X_i, X_j, \mathbf{H}) = \hat{T}(X_i, X_j|\mathbf{Pa}_j^{\mathcal{G}})$.
    if $Score(X_i, X_j, \mathbf{H}) < s$ then
        Let $s = Score(X_i, X_j, \mathbf{H})$ and $I = Delete(X_i, X_j, \mathbf{H})$.
    end if
end for
if $s < \tau$ then
    Apply operator $I$ to obtain $\mathcal{P}'$.
else
    Keep $\mathcal{P}' = \mathcal{P}$ (and terminate BES).
end if
return $\mathcal{P}'$
```

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 $(0, m_0)$.

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

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 $\mathbf{D} = \{(x_i, y_i, z_i), i = 1, \dots, n\}$, 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

$$\begin{aligned}&\sup_{\theta\in\Theta} \left|\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\hat{\omega}}(Z)]^2 - \mathbb{E}[f_\theta(X,Z) - h_{\omega^*}(Z)]^2\right| \\ \le &\sup_{\theta\in\Theta} \left|\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\hat{\omega}}(Z)]^2 - \hat{\mathbb{E}}[f_\theta(X,Z) - h_{\omega^*}(Z)]^2\right| + \sup_{\theta\in\Theta} \left|\hat{\mathbb{E}}[f_\theta(X,Z) - h_{\omega^*}(Z)]^2 - \mathbb{E}[f_\theta(X,Z) - h_{\omega^*}(Z)]^2\right|\end{aligned} \tag{1}$$

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:

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

We recall the definitions

$$\omega^*(\theta) = \operatorname*{argmin}_{\omega\in\Omega} \mathbb{E}[f_\theta(X,Z) - h_\omega(Z)]^2$$

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

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 $n \to \infty$,

$$\sup_{\theta\in\Theta, z\in K} |h^2_{\hat{\omega}(\theta)}(z) - h^2_{\omega^*(\theta)}(z)| \xrightarrow{p} 0. \tag{3}$$

Next, we show the second term in (2) vanishes in probability. Given an arbitrary $r > 0$, let $B_r = \{z \in \mathbb{R}^{dz} : \|z\| \le r\}$. Let $B_r^c = \mathbb{R}^{dz} \setminus B_r$ be its complement. We have

$$\begin{aligned}
\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left[h^2_{\hat{\omega}}(z_i) - h^2_{\omega^*}(z_i)\right] \right| &\le \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \left|h^2_{\hat{\omega}}(z_i) - h^2_{\omega^*}(z_i)\right| \\
&= \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \left[\left|h^2_{\hat{\omega}}(z_i) - h^2_{\omega^*}(z_i)\right| \mathbf{1}_{\{z_i\in B_r\}} + \left|h^2_{\hat{\omega}}(z_i) - h^2_{\omega^*}(z_i)\right| \mathbf{1}_{\{z_i\in B_r^c\}}\right] \\
&\le \sup_{\theta\in\Theta, z\in B_r} |h^2_{\hat{\omega}}(z) - h^2_{\omega^*}(z)| + \frac{2}{n}\sum_{i=1}^n H^2(z_i)\mathbf{1}_{\{z_i\in B_r^c\}},
\end{aligned} \tag{4}$$

where the second term in the upper bound (4) comes from the dominated integrable condition in *C3* with a dominating function $H(z)$. By taking $n \to \infty$, the first term in (4) vanishes in probability by (3), and the second term in (4) 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 $n \to \infty$

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left[h^2_{\hat{\omega}}(z_i) - h^2_{\omega^*}(z_i)\right] \right| \xrightarrow{p} 0. \tag{5}$$

Last, we show the first term in (2) vanishes in probability. Again, we consider an arbitrary radius $r > 0$ and a compact ball $B'_r = \{(x,z) \in \mathbb{R}^{dx+dz} : \|(x,z)\| \le r\}$. 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

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

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 $n \to \infty$

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n f_\theta(x_i,z_i)\left[h_{\hat{\omega}}(z_i) - h_{\omega^*}(z_i)\right] \right| \xrightarrow{p} 0. \tag{6}$$

Then by combining convergence results (5) and (6) and recalling the upper bounds (1) and (2), we have as $n \to \infty$, (ii) holds. Similarly we can show (iii). □

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

$$\begin{aligned}
&\sup_{\theta\in\Theta,\phi\in\Phi} \left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\hat{\omega}}(Z))\cdot(g_\phi(Y,Z) - l_{\hat{\psi}}(Z))] - \mathbb{E}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right| \\
\le& \sup_{\theta\in\Theta,\phi\in\Phi} \left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\hat{\omega}}(Z))\cdot(g_\phi(Y,Z) - l_{\hat{\psi}}(Z))] - \hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right| \\
&+ \sup_{\theta\in\Theta,\phi\in\Phi} \left|\hat{\mathbb{E}}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))] - \mathbb{E}[(f_\theta(X,Z) - h_{\omega^*}(Z))\cdot(g_\phi(Y,Z) - l_{\psi^*}(Z))]\right|,
\end{aligned} \tag{7}$$

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 (7) is upper bounded by

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

where all three terms converge to 0 in probability as $n\to\infty$. Therefore, the left-hand side of (7) 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

$$\begin{aligned}\mathbb{E}[f(X,Z)-h(Z)]^2 &= \mathbb{E}[(f(X,Z)-h^*(Z))+(h^*(Z)-h(Z))]^2\\ &= \mathbb{E}[f(X,Z)-h^*(Z))]^2+\mathbb{E}[h^*(Z)-h(Z)]^2+\mathbb{E}[(f(X,Z)-h^*(Z))(h^*(Z)-h(Z))].\end{aligned}\tag{8}$$

Note that the cross term in the second line of (8) 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 (8) 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 $\mathbb{R}^{dz}$. Let $M(i)$ be the index j such that (X_j,Z_j) is the nearest neighbor of (X_i,Z_i) in $\mathbb{R}^{dx+dz}$. 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 [Ramsey et al., 2017] 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 1; 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 EXPERIMENTAL DETAILS AND HYPERPARAMETERS

Since our model is based on deep neural networks (NNs), 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 turn out to be very robust across different settings so we keep them 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.

Roughly speaking, as we have more nodes and edges, we need larger NNs with more layers and neurons per layer. We suggest that practitioners tune the architecture on synthetic data with the same number of nodes and edges (roughly) and transfer the hyperparameters to the datasets at hand. The global score proposed below in (9) can serve as a metric

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 1: Hyperparameters of NCD and RND for all settings.

to evaluate each set of hyper-parameter values: a good set of hyper-parameter values should be the one that yields a low score (approaching 0) for the true structure and higher scores for any fake structures. In our experiments, we tune the hyperparameters on synthetic data sampled from additive noise models and transfer them to the PNL datasets, etc. Moreover, when some prior knowledge on the ground-truth structure is available, such as the absence or presence of a few edges and their orientations (which may imply some conditional independence conditions), we suggest tuning the hyper-parameters to match the prior information as much as possible.

We also proposed in the paper an implementation of the reframed GES with the RCD measure in the literature, which involves no NN hyper-parameters and performs reasonably well across various settings. The only hyperparameter of the RCD implementation is the threshold τ. In other words, when using the reframed GES in practice, there is also a good option that does not involve much hyper-parameter tuning, with some loss of accuracy in certain settings in comparison to the NN implementation NCD. We listed the specific hyperparameters for the our experiments in Table 1.

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 $\mathcal{G}$:

$$S_g(\mathcal{G}) = \frac{1}{d}\sum_{i=1}^{d} \hat{S}_n(X_i, \mathbf{Nd}_i^{\mathcal{G}} | \mathbf{Pa}_i^{\mathcal{G}}). \tag{9}$$

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 2-3 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 is 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.

Moreover, we present the results of all methods on the PNL data sets with 20 nodes, 2 expected degrees and 5000 samples in Tables 4-5. We observe that our reframed GES with NCD or RCD performs among the best methods in this setting.

In addition to the performance in causal discovery, we also compare the computational time of different methods. We consider four methods related to the GES algorithm: the standard GES with the BIC score (BIC), the standard GES with GSF score (GSF), our reframed GES with the RCD and NCD score. Table 6 reports the average running time of each method. As mentioned in the main text, our proposed NCD measure has an advantage over the kernel-based GSF in computation, both of which are nonparametric causal discovery methods. It has been well acknowledged that kernel methods suffer from high sample complexity (although some more efficient approximations exist), while neural networks can benefit from a large sample size without a severe compromise in computational time. We see from Table 6 that NCD is significantly more computationally efficient than GSF, especially on datasets with a sparse graph structure. Moreover, since the NCD estimator is obtained by applying Algorithm 2, it is much more computationally demanding than score functions with an explicit formula to be easily computed such as BIC and RCD. This is clearly verified by the results in Table 6.

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 2: 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 3: 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 4: 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 5: Results on PNL-MULT data with 20 nodes, 2 expected degrees and 5000 samples.

Dataset	Node	Degree	BIC	RCD	NCD	GSF
PNL-GP	10	2	< 1	< 1	54.8	1195.2
PNL-MULT	10	2	< 1	1.2	305	2122.8
PNL-GP	10	8	< 1	3.8	748.8	1801.2
PNL-MULT	10	8	< 1	9.2	618.0	1854.0
PNL-GP	20	2	< 1	2.0	253.2	6461.4
PNL-MULT	20	2	< 1	3.66	194.4	6380.4

Table 6: Average running time (seconds).

RCD, as a nonparametric measure, costs more to compute than the simple BIC score, but is still fairly fast compared with the other two.

In fact, it is a trade-off between computational complexity and the quality of statistical estimation. As we noted in the main text, BIC is consistent only in restrictive parametric cases; otherwise, there exists a systematic error due to model misspecification, leading usually to poor results. This is verified by extensive results in causal discovery shown above and in the main text. In contrast, our NCD estimator is consistent in nonparametric settings which is much more general and flexible. Therefore, in applications of causal discovery where the accuracy of the estimation matters more than the computational cost, our approach has advantages over BIC.

References

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070 – 3102, 2021.

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

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

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-`.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.