CAUSAL DISCOVERY FOR LINEAR CAUSAL MODEL WITH CORRELATED NOISE: AN ADVERSARIAL LEARNING APPROACH

Mujin Zhou
Tsinghua University

Junzhe Zhang
Syracuse University
jzhan403@syr.edu

ABSTRACT

Causal discovery from data with unmeasured confounding factors is a challenging problem. This paper proposes an approach based on the f-GAN framework, learning the binary causal structure independent of specific weight values. We reformulate the structure learning problem as minimizing Bayesian free energy and prove that this problem is equivalent to minimizing the f-divergence between the true data distribution and the model-generated distribution. Using the f-GAN framework, we transform this objective into a min-max adversarial optimization problem. We implement the gradient search in the discrete graph space using Gumbel-Softmax relaxation.

***K*eywords** Causal Discovery · Unmeasured Confounding · f-GAN

1 Introduction

Causal discovery infers potential causal relationships between variables from observed data, and is therefore applied in fields such as biology, econometrics, and social sciences. Directed acyclic graphs (DAGs) are classic models used to describe causal relationships between observable variables. However, DAGs are insufficient to describe the unmeasured confounding that is prevalent in complex real-world systems. More general graph structures are needed, such as explicitly modeling the error correlations caused by confounding by introducing bi-directed edges.

To learn these structures from data, differentiable score-based continuous optimization methods have made significant progress in recent years. Bhattacharya et al. (2021) [1] extended continuous optimization methods for DAGs to linear systems with unmeasured confounding. This method derives a series of differentiable algebraic constraints to characterize specific ADMG subclasses, particularly Arid and Bow-free ADMGs. Under these specific graph class constraints, Bhattacharya et al. proved that the model parameters are identifiable, meaning the model belongs to the regular statistical class. Therefore, they used the approximate BIC as the scoring function and obtained the optimal parameter point estimates that satisfy the constraints through the augmented Lagrangian method. However, if model selection is desired for more general model classes, without imposing restrictions such as Arid or Bow-free (i.e., allowing more general bidirectional edge structures), causal models often exhibit singularity, rendering BIC no longer an effective approximation of model evidence.

Watanabe (2009)'s singular learning theory states that for regular models, the Bayes free energy can be effectively approximated by BIC. However, for singular models, the likelihood function cannot be approximated by any normal distribution, causing BIC to fail. In this context, the model selection criterion can be directly based on Bayesian free energy, defined as $F = -\log \int \prod p(X_i|w)\phi(w)dw$, where $\phi(w)$ is the prior distribution of the parameters[2].

Inspired by this theory, this paper proposes a generative framework-based causal structure learning method to estimate the general causal graph (DAG) structure with unmeasured confounding from observed data. This problem is constructed as a model selection problem, aiming to infer a binary causal structure M rather than specific weight parameters. Each model M is defined by an adjacency matrix S_B representing direct causal relationships and a noisy covariance structure S_Σ representing the correlations between variables caused by unmeasured confounding, i.e., $M = (S_B, S_\Sigma)$.

This paper employs Bayesian free energy as the criterion for model selection and derives that minimizing the expected free energy of the data is equivalent to minimizing the Kullback-Leibler divergence between the true data distribution

P_{data} and the marginal likelihood $q(\cdot|M)$. Then, using the f-GAN framework [3], where KL divergence is a special case of f-divergence, this problem is transformed into a training task for a Generative Adversarial Network (GAN). To address the issue that the causal graph structures S_B and S_Σ are discrete and cannot be directly optimized using gradient descent, this paper uses the Gumbel-Softmax relaxation technique to transform the discrete structure selection problem into an optimization problem for continuous parameters. This allows the entire model to be trained through gradient backpropagation and ultimately learn the graph structure.

2 Preliminaries

This section first defines the mathematical notation for linear structural equation models with unmeasured confounding. It then reviews the definition of f-divergence and details the f-GAN Variational Divergence Minimization framework proposed by Nowozin et al.[3], which forms the theoretical basis of the method presented in this paper.

2.1 Linear SEM with unmeasured confounding

Consider a system with d observed variables, denoted by a random vector $X = [X_1, \ldots, X_d] \in \mathbb{R}^d$. Assume the data generation process follows a Linear Structural Equation Model (SEM):

$$X = XB + E \tag{1}$$

where $B \in \mathbb{R}^{d\times d}$ is a weighted adjacency matrix, and $B_{ij} \neq 0$ indicates a direct causal effect from variable X_i to X_j. $E = [E_1, \ldots, E_d] \in \mathbb{R}^d$ is a noise vector, assumed to follow a multivariate Gaussian distribution with a mean of zero $E \sim \mathcal{N}(0, \Sigma)$.

Since matrix B corresponds to a DAG, matrix $(I - B)$ is invertible. Therefore, the generation process of observed data X is:

$$X = E(I - B)^{-1} \tag{2}$$

The goal of this paper is not to estimate the specific parameter matrices B and Σ, but to recover their underlying binary structures, defined as $M = (S_B, S_\Sigma)$:

- $S_B \in \{0, 1\}^{d\times d}$: represents the adjacency matrix of the DAG, where $(S_B)_{ij} = 1 \iff B_{ij} \neq 0$, indicating the existence of a directed edge from X_i to X_j.
- $S_\Sigma \in \{0, 1\}^{d\times d}$: Represents a confounding correlation structure, where $(S_\Sigma)_{ij} = 1 \iff \Sigma_{ij} \neq 0$ and $i \neq j$, indicating the existence of a bidirectional edge connecting X_i and X_j, meaning there exists a latent variable that simultaneously affects both X_i and X_j.

2.2 Variational Divergence Minimization (f-GAN)

One step in our approach to learning the causal structures is to estimate the f-divergence (defined later) between the true data distribution and the model distribution. We first review the main theoretical foundation of our method, namely the f-GAN framework, which is based on and extends the method for estimating f-divergence proposed by Nguyen et al. (2010)[4]. Through a process called Variational Divergence Minimization (VDM), f-GAN transforms the task of minimizing f-divergence into an optimization problem of generative adversarial networks.

Definition of f-divergence Given two probability distributions P and Q, with probability density functions $p(x)$ and $q(x)$ respectively, defined on the domain $\mathcal{X}$. f-divergence is defined as:

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{3}$$

Where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function and satisfies $f(1) = 0$. Different generator functions f correspond to different divergences, for example, $f(u) = u \log u$ corresponds to Kullback-Leibler (KL) divergence, and $f(u) = -\log u$ corresponds to Reverse KL divergence.

Variational Lower Bound Because the analytical form of the true distribution is unknown, f-divergence cannot be directly calculated. Nguyen et al. (2010) derived a variational lower bound for the f-divergence using the Fenchel conjugate.

The Fenchel conjugate f^* of a function f is defined as:

$$f^*(t) = \sup_{u \in \text{dom}_f} \{ut - f(u)\} \tag{4}$$

Since f^* is again a convex function and lower semi-continuous, $f^{**} = f$. We can express $f(u)$ as $f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\}$. Substituting this into the definition of $D_f(P\|Q)$ in (3), and letting $u = \frac{p(x)}{q(x)}$, we get:

$$\begin{aligned} D_f(P\|Q) &= \int q(x) \sup_t \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx \\ &\geq \sup_{T \in \mathcal{T}} \left(\int q(x) \cdot T(x) \frac{p(x)}{q(x)} dx - \int q(x) f^*(T(x)) dx \right) \\ &= \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[f^*(T(x))] \right) \end{aligned} \tag{5}$$

Where $T : \mathcal{X} \to \mathbb{R}$ is any class of function (parameterized by the discriminator network in f-GAN). The inequality is derived from the Jensen inequality and the exchange of integral and supremum operations. This lower bound is tight when the class $\mathcal{T}$ is sufficiently large.

Generative Adversarial Training Objective: In GAN training, we aim to find a set of parameters θ such that the model distribution Q_θ approximates the real data distribution P_{data} as closely as possible, i.e., $\min_\theta D_f(P_{data}\|Q_\theta)$. Combining the variational lower bound of $D_f(P_{data}\|Q_\theta)$ in (5), this problem can be transformed into solving the following Min-Max objective:

$$\min_\theta \max_\omega V(G_\theta, T_\omega) = \mathbb{E}_{x \sim P_{data}}[T_\omega(x)] - \mathbb{E}_{x \sim Q_\theta}[f^*(T_\omega(x))] \tag{6}$$

Here, G_θ represents the generator, and T_ω represents the variational function (i.e., the discriminator). This framework unifies the general f-divergence minimization problem into the adversarial training problem.

3 Method

Based on the previous theoretical foundation, this section elaborates on our proposed causal structure learning method. The core idea is to treat the causal structure as a model and use Bayesian free energy as the criterion for model selection. Then, it is derived that minimizing Bayesian free energy is equivalent to minimizing the divergence between distributions, and this can be solved within the f-GAN framework.

3.1 Bayesian Free Energy and Model Selection

The goal is to select the optimal binary causal structure $M = (S_B, S_\Sigma)$ from the observed data $\mathbf{X}$, rather than to estimate the values of specific parameters. From a Bayesian perspective, the model evidence, or marginal likelihood, for a given structure M is obtained by integrating over all possible weight parameters $w = (B, \Sigma)$:

$$q(\mathbf{X}|M) = \int q(\mathbf{X}|w, M)\phi(w|M)dw = \int \left(\prod_{i=1}^{n} q(\mathbf{X}_i|w, M) \right) \phi(w|M)dw \tag{7}$$

where $\phi(w|M)$ is the prior distribution of the parameters given the structure M (e.g., a uniform distribution at the locations where edges exist).

The Bayes Free Energy of M, $F(\mathbf{X}|M)$, is defined as the negative logarithm of the model's evidence:

$$F(\mathbf{X}|M) = -\log q(\mathbf{X}|M) \tag{8}$$

For singular models containing latent variables or hierarchical structures, F cannot be asymptotically approximated by BIC. So the Bayes Free Energy is directly used as the criterion for model selection. Our goal is to find a structure $M = (S_B, S_\Sigma)$ such that the expected value of this free energy is minimized under the true data distribution P_{data}, with the corresponding directed graph being acyclic:

$$\begin{aligned} &\min_{M} \mathbb{E}_{x\sim P_{data}}[F(\mathbf{X}|M)] \\ &\text{s.t. } S_B \in \text{DAGs} \end{aligned} \tag{9}$$

It can be proved that this objective is equivalent to minimizing the Kullback-Leibler (KL) divergence between the true distribution P_{data} and the model distribution $q(X|M)$. By adding and subtracting the entropy of the true data distribution, $H(\mathbf{X}) = -\int P_{data}(\mathbf{X}) \log P_{data}(\mathbf{X}) d\mathbf{X}$:

$$\begin{aligned} \mathbb{E}_{x\sim P_{data}}[F(x|M)] &= -\int P_{data}(\mathbf{X}) \log q(\mathbf{X}|M) d\mathbf{X} \\ &= \int P_{data}(\mathbf{X}) \log \frac{P_{data}(\mathbf{X})}{q(\mathbf{X}|M)} d\mathbf{X} + H(\mathbf{X}) \\ &= D_{KL}(P_{data}||q(\cdot|M)) + H(\mathbf{X}) \end{aligned} \tag{10}$$

Since $H(\mathbf{X})$ is a constant with respect to the model M:

$$\arg\min_{M} \mathbb{E}_{x\sim P_{data}}[F(\mathbf{X}|M)] \iff \arg\min_{M} D_{KL}(P_{data}||q(\cdot|M))$$

Now the minimization problem turns into the minimization of the KL divergence between the true data distribution and the model distribution:

$$\begin{aligned} &\arg\min_{M} D_{KL}(P_{data}||q(\cdot|M)) \\ &\text{s.t. } S_B \in \text{DAGs} \end{aligned} \tag{11}$$

3.2 Optimization via f-GAN Framework

Directly calculating the KL divergence requires high-dimensional integrals. Using the f-GAN framework introduced in the Preliminaries, we can minimize the lower bound of this divergence through adversarial training.

Although the theoretical derivation points to the KL divergence (corresponding to $f(u) = u \log u$), we adopt the objective of the original GAN for convenience. The standard GAN objective is equivalent to minimizing the Jensen-Shannon (JS) divergence.

Deriving the Standard GAN Objective Function from f-GAN: As shown by Nowozin et al. (2016)[3], the generator function f corresponding to the standard GAN objective and its Fenchel conjugate f^* are, respectively:

$$\begin{aligned} f(u) &= u \log u - (u+1) \log(u+1) \\ f^*(t) &= -\log(1 - e^t) \end{aligned} \tag{12}$$

According to the variational lower bound formula in Eq. (5):

$$D_f(P\|Q) \geq \sup_{T\in\mathcal{T}} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \tag{13}$$

The model distribution is generated by a generator G parameterized by $M = (S_B, S_\Sigma)$, so we rewrite the marginal distribution $q(\cdot|M)$ as $Q_G(S_B, S_\Sigma)$. We use a neural network discriminator $D_\omega(x)$ to serve as the variational function $T(x)$. The mapping between the variational function $T(x)$ and the discriminator output $D_\omega(x)$ is designated as:

$$T(x) = \log D_\omega(x) \tag{14}$$

Substituting Eq.(14) and $f^*(t)$ into the variational lower bound formula:

$$
\begin{aligned}
D_f(P\|Q) &\geq \sup_{T\in\mathcal{T}} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \\
&= \max_{\omega} \mathbb{E}_{x\sim P}[\log D_\omega(x)] - \mathbb{E}_{x\sim Q}[f^*(\log D_\omega(x))] \\
&= \max_{\omega} \mathbb{E}_{x\sim P}[\log D_\omega(x)] - \mathbb{E}_{x\sim Q}[-\log(1-e^{\log D_\omega(x)})] \\
&= \max_{\omega} \mathbb{E}_{x\sim P}[\log D_\omega(x)] + \mathbb{E}_{x\sim Q}[\log(1-D_\omega(x))]
\end{aligned} \tag{15}
$$

Thus the standard GAN's Min-Max objective is recovered. Substituting (15) into the objective of $\min D_f(P\|Q)$, we obtain $V(G,D)$:

$$
\min_G \max_{D_\omega} V(G,D) = \mathbb{E}_{x\sim P_{data}}[\log D_\omega(x)] + \mathbb{E}_{x\sim Q_G}[\log(1-D_\omega(x))] \tag{16}
$$

This derivation shows that by optimizing the objective function of a standard GAN, we are actually variationally minimizing the f-divergence between the true distribution and the model's marginal distributions, thereby indirectly optimizing the Bayes free energy.

3.3 Generative Process

For the above framework to work, the generator G must be able to sample from the marginal distribution $Q(X|M)$. Since the binary structures S_B, S_Σ are discrete and cannot be directly used as parameters for G in differentiable optimization, the generator instead uses two logit matrices as parameters: $A_B \in \mathbb{R}^{d\times d}$ and $A_\Sigma \in \mathbb{R}^{d\times d}$, corresponding to the existence probabilities of directed and bidirectional edges, respectively. The process of generating a single sample X_{fake} is as follows:

1. Structure Sampling: Generate soft approximations of the binary structure based on logits, denoted as $\tilde{S}_B$ and $\tilde{S}_\Sigma$ (see the next section for specific implementation of the Gumbel-Softmax trick).

2. Prior Sampling: Randomly sample the specific weight matrix from the predefined prior distribution $\phi(w)$. For example, in this paper it is sampled from a uniform distribution: $B' \sim \mathcal{U}([-2,-0.5]\cup[0.5,2])$

3. Masking: Combining the sampled structure with random weights to obtain the instance parameters generated:

$$
B^* = \tilde{S}_B \odot B', \quad \Sigma^* = \tilde{S}_\Sigma \odot \Sigma' \tag{17}
$$

4. Data Generation: Generating data based on a linear SEM:

$$
E \sim \mathcal{N}(0,\Sigma^*), \quad X_{fake} = E(I-B^*)^{-1} \tag{18}
$$

The generation process involves sampling given a structure M and weights w, and marginalizes the parameters through random sampling of the weights. The samples generated in each batch follow the distribution $q(\mathbf{X}|M) = \int \left(\prod_{i=1}^k q(\mathbf{X}_i|w,M)\right)\phi(w|M)dw$.

3.4 Differentiable Relaxation and Constraints

To achieve gradient descent optimization, we need to handle the non-differentiable discrete structure and the acyclic constraint of the DAG.

Gumbel-Softmax Relaxation: We apply the Gumbel-Softmax trick[5] to transform discrete sampling into a differentiable operation. For each potential edge (i,j), the possibility of the binary choice between "no edge" (0) and "edge" (1) is parameterized by the logits $[0, A_{B,ij}]$. To obtain the discrete structure S_B, ideally, sampling should be performed at each data generation stage from a Bernoulli distribution with probabilities $P_{B,ij} = \text{sigmoid}(A_{B,ij})$. However, since the discrete sampling process is non-differentiable, the gradient cannot be backpropagated to the logit matrix A_B. The Gumbel-Softmax provides a continuous and differentiable sample from the corresponding categorical distribution:

$$
(\tilde{S}_B)_{ij} = \text{softmax}\left(\frac{[g_0, A_{B,ij}+g_1]}{\tau}\right)_1 \tag{19}
$$

where $g_0, g_1 \sim \text{Gumbel}(0,1)$ are independent noise samples and τ is a temperature parameter. As $\tau \to 0$, $\tilde{S}_B$ approaches discrete $\{0,1\}$ values. The resulting matrix, $\tilde{S}_B$, serves as a differentiable proxy for the binary structure S_B.

Algorithm 1 Generative Causal Discovery via f-GAN

Require: Observed data $\mathbf{X}$; Acyclic constraint parameter λ_{acyc}; Learning rates η_D, η_G; Batch size k; Epoch T; Gumbel-Softmax temperature τ_{start}, τ_{end} and its anneal rate.

Initialize: Discriminator parameters ω; Generator parameters A_B, A_Σ.

procedure GENERATOR(Z, A_B, A_Σ, τ)
  $\tilde{S}_B, \tilde{S}_\Sigma \leftarrow$ GumbelSoftmax(A_B, A_Σ, τ)
  Randomly sample weights matrices $B' \in \mathbb{R}^{p\times p}, \Sigma' \in \mathbb{R}^{p\times p}$
  $B^* \leftarrow \tilde{S}_B \odot B'$, $\Sigma^* \leftarrow \tilde{S}_\Sigma \odot \Sigma'$
  $L^* \leftarrow$ Cholesky(Σ^*)
  $X_{fake} \leftarrow Z{L^*}^T(I - B^*)^{-1}$
  return $X_{fake}, \tilde{S}_B$
end procedure

for epoch $= 1, 2, \ldots, T$ **do**
  for each batch of k real samples X_{real} from $\mathbf{X}$ **do**
    // — Train Generator G — //
    Sample noises: Batch $Z \in \mathbb{R}^{k\times p} \sim \mathcal{N}(0, I)$
    $X_{fake}, \tilde{S}_B \leftarrow$ GENERATOR(Z, A_B, A_Σ, τ)
    $L_G \leftarrow L_{adv} + \lambda_{acyc} \cdot \text{tr}(e^{\tilde{S}_B \circ \tilde{S}_B} - p)$
    $A_B \leftarrow A_B - \eta_G \nabla_{A_B} L_G$
    $A_\Sigma \leftarrow A_\Sigma - \eta_G \nabla_{A_\Sigma} L_G$
    // — Train Discriminator D — //
    $X_{fake} \leftarrow$ GENERATOR(Z, A_B, A_Σ, τ)
    $\omega \leftarrow \omega - \eta_D \nabla_\omega L_D$
  end for
  Anneal Gumbel-Softmax temperature τ
end for

// — Final Structure Extraction — //
$P_B \leftarrow \text{sigmoid}(A_B), \quad P_\Sigma \leftarrow \text{sigmoid}(A_\Sigma)$
$S_B^* \leftarrow (P_B > \delta), \quad S_\Sigma^* \leftarrow (P_\Sigma > \delta)$
return S_B^*, S_Σ^*

A similar procedure is applied to A_Σ to obtain $\tilde{S}_\Sigma$, the differentiable proxy for S_Σ. The acyclic constraint $h(S_B)$ is also substituted by $h(\tilde{S}_B)$.

3.5 Algorithm Summary

With the differentiable generator defined, the precise loss functions used to train the discriminator and the generator can be specified, based on the min-max objective $V(G, D)$.

- Discriminator Loss:

$$L_D = -\mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] - \mathbb{E}_{Z\sim\mathcal{N}(0,I)}[\log(1 - D(G(Z; A_B, A_\Sigma)))]$$

- Generator Loss (In practice, the adversarial component $\mathbb{E}_Z[1 - \log(D(x))]$ is replaced with $-\mathbb{E}_Z[\log(D(x))]$):

$$L_G = -\mathbb{E}_{Z\sim\mathcal{N}(0,I)}[\log(D(G(Z; A_B, A_\Sigma)))] + \lambda_{\text{acyc}} \cdot h(\tilde{S}_B)$$

After training, we extract the final causal structure by applying the Sigmoid function to Logits A_B, A_Σ and setting a threshold.

The complete procedure is summarized in Algorithm 1.

4 Experiments

In this section, we evaluate the performance of our proposed method, **fGAN-CD** (f-GAN Causal Discovery), on synthetic datasets generated from linear Structural Equation Models (SEMs) with unmeasured confounding. We compare our approach against the ABIC bow-free method [1].

4.1 Experimental Setup

Data Generation. We generate synthetic datasets based on linear SEMs with correlated errors: $X = E(I - B)^{-1}$, where $E \sim \mathcal{N}(0, \Sigma)$. The covariance matrix Σ is non-diagonal, simulating the presence of unmeasured confounders. We adopt a setting similar to [1]:

- **Structure:** We consider two fixed ground truth binary graphs $G^* = (S_B, S_\Sigma)$.
- **Edge Weights (B):** For each directed edge (i, j) in S_B, the weight B_{ij} is sampled from $\mathcal{U}([-2.0, -0.5] \cup [0.5, 2.0])$.
- **Noise Covariance (Σ):** Diagonal elements (variances) are sampled from $\mathcal{U}([0.7, 1.2])$. Off-diagonal elements corresponding to bidirected edges in S_Σ are sampled from $\mathcal{U}([-0.7, -0.4] \cup [0.4, 0.7])$.

For each case study, we generate $N = 2000$ i.i.d. samples.

Baselines & Implementation. We compare our method against **ABIC** (Differentiable Causal Discovery under Unmeasured Confounding) [1]. For ABIC, we use the official implementation with default hyperparameters ($\lambda = 0.05$). For our fGAN-CD, we use a Gumbel-Softmax temperature decay from 1.0 to 0.1 over 4000 epochs.

Evaluation Metrics. Since different ADMGs can be Markov equivalent, we convert both the ground truth ADMG and the estimated graphs into their corresponding **Partial Ancestral Graphs (PAGs)**, which uniquely represent the Markov equivalence class. We report the following metrics on the PAGs:

- **SHD (Structural Hamming Distance)**: The count of all edge additions, removals, and direction changes needed to convert the estimated PAG into the true PAG.
- **Skeleton F1**: The F1 score of the existence of edges, ignoring orientation.
- **Arrowhead F1**: The F1 score of correctly recovering arrowheads ($\rightarrow$).

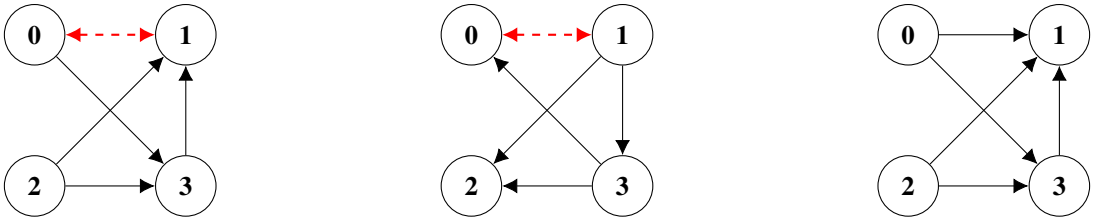
4.2 Case Study A: Structure Recovery

In this experiment, the ground truth structure consists of 4 nodes with directed edges $\{(2, 1), (2, 3), (0, 3), (3, 1)\}$ and a bidirected edge $\{(0, 1)\}$. We generated $N = 2000$ samples.

Table 1: Quantitative comparison on Case Study A ($N = 2000$). Best results are bolded.

Method	SHD ↓	Skeleton F1 ↑	Arrowhead F1 ↑
ABIC	3.4	0.897	0.1
fGAN-CD	2.17	0.909	0.667

Table 1 summarizes the results, Figure 1 visualizes the true structure and the structure learned by methods.



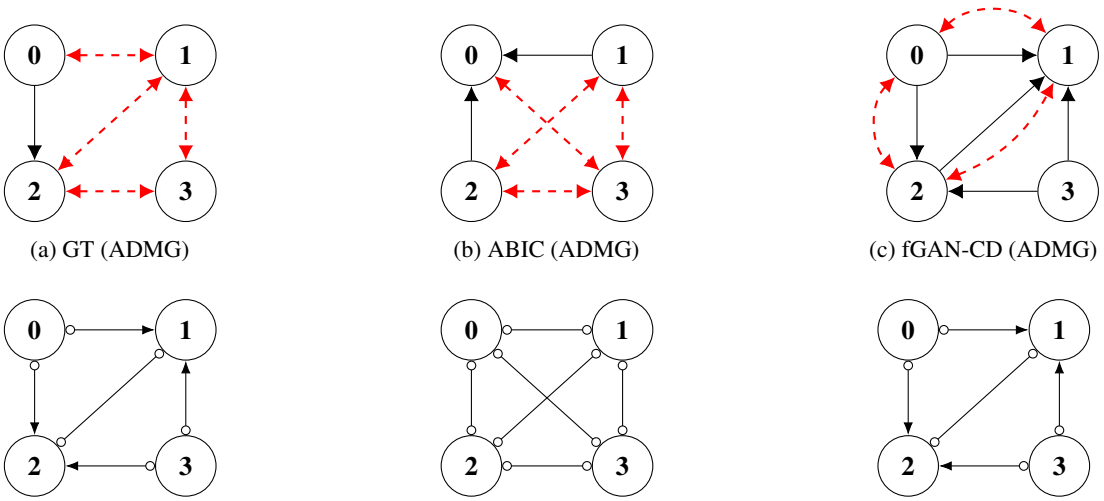
(a) Ground Truth (b) ABIC (Baseline) (c) fGAN-CD

Figure 1: Comparison of learned structures. Red dashed lines represent bidirected edges (confounders). It can be seen that fGAN-CD has a greater advantage in learning the arrow directions.

4.3 Case Study B: Sparsity and Independence Test

An important objective of causal discovery is to recover the conditional independence constraints in the true graph as much as possible. To evaluate this, we designed a high confounding scenario with ground truth: $D = \{(0, 2)\}$ and a dense confounding structure $B = \{(0, 1), (1, 2), (1, 3), (2, 3)\}$.

A key property of this graph is that Node 0 and Node 3 are independent, despite being connected to common neighbors via bidirected edges. Path $0 \leftrightarrow 1 \leftrightarrow 3$: Node 1 is a collider; Path $0 \rightarrow 2 \leftrightarrow 3$: Node 2 is a collider; Path $0 \rightarrow 2 \leftrightarrow 1 \leftrightarrow 3$: Nodes 1 and 2 are both colliders. Each path is blocked. So one metric for measuring the capability of a causal discovery method is whether it can indicate no edge between 0 and 3 in the resulting PAG. We use this as a sanity check.



(a) GT (ADMG) (b) ABIC (ADMG) (c) fGAN-CD (ADMG)

(d) GT (PAG) (e) ABIC (PAG) (f) fGAN-CD (PAG)

Figure 2: Comparison of Learned Structures (Top Row) and their corresponding PAGs (Bottom Row). In the ADMGs, red dashed lines indicate unmeasured confounding.

As shown in Figure 2, ABIC incorrectly inferred a connection between nodes 0 and 3. The fGAN-CD method, however, correctly identifies that no edge exists between nodes 0 and 3 and recovers the true PAG. Furthermore, although we set up a bow-free ground truth graph to accommodate the constraints of the ABIC method, fGAN-CD can actually overcome these contraints and learn arbitrary ADMG graphs.

References

[1] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2314–2322. PMLR, 2021.

[2] Sumio Watanabe. A widely applicable bayesian information criterion, 2012.

[3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization, 2016.

[4] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, November 2010.

[5] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.