Boosting Causal Structure Learning: An Asymmetric Exponential Modulation Gaussian-Based Adaptive Sample Reweighting Framework

Wei Xiao, Hongbin Wang*, Ming He, Nianbin Wang

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
{_xiaowei_, wanghongbin, heming, wangnianbin}@hrbeu.edu.cn

Abstract

Recent advances in differentiable score-based methods for Directed Acyclic Graph (DAG) structure learning have revolutionized the problem of combinatorial structure learning, transforming it into a continuous optimization task. Despite their remarkable success, these methods rely on a key assumption that all samples have the same level of difficulty and no data heterogeneity. When this assumption does not hold, causal discovery algorithms based on it inevitably return networks with many spurious edges. Despite existing research, the current method ignores the reality of outliers in the samples, introducing certain limitations that still result in erroneous edges. Inspired by the rapid decay of the Gaussian distribution as distance from the center increases, we propose an innovative adaptive sample reweighting framework based on asymmetric exponential modulation Gaussian, coined DAG-AEG. DAG-AEG boosts DAG structure learning by analyzing the distribution of sample losses and employing the proposed method for adaptive sample attention. Additionally, it can be adapted to heterogeneous data. We used various causal structure learning methods to test the performance of DAG-AEG on synthetic and real datasets. The experimental results demonstrate that the proposed framework significantly improves the performance across all methods, outperforming existing methods.

Introduction

Causal structure learning aims to identify causal relationships between variables from observational data, serving as a foundation for research in many scientific fields (Pearl 2009; Liang et al. 2024). It has significant applications across various domains(Sachs et al. 2005; Pearl 2019; Locatello et al. 2019; Castro, Walker, and Glocker 2020). Additionally, the discovery of causal relationships is considered one of the essential tools for advancing from current Artificial Narrow Intelligence to Artificial General Intelligence (Pearl 2018).

An effective method for discovering causal relationships between variables is to conduct randomized experiments (Boruch 1997). However, in practical applications, this can be quite costly and may even be prohibited in some cases due to ethical concerns. Therefore, learning causal structures from purely observational data has become a research hotspot in recent years (Ng et al. 2019; Xu et al. 2020; Li et al. 2020).

Current methods for causal structure learning can be divided into constraint-based and score-based approaches (Vowels, Camgoz, and Bowden 2022). Constraint-based methods, such as the PC (Spirtes, Glymour, and Scheines 2001) and FCI (Spirtes, Meek, and Richardson 2013) algorithms, use conditional independence tests and a series of rules to identify causal directions. Score-based methods leverage predefined score functions to evaluate all candidate graphs in the DAG space to find the optimal causal graph. However, due to the combinatorial acyclicity constraint of causal graphs, finding the score-optimal causal graph is usually NP-hard (Chickering, Heckerman, and Meek 2004). Recently, (Zheng et al. 2018) proposed a new smooth acyclicity constraint, transforming the combinatorial optimization problem into a continuous one, making it possible to optimize the score function through gradient descent. Subsequent differentiable causal discovery methods have extended this approach to nonlinear problems by utilizing various neural network models (Yu et al. 2019; Zhu, Ng, and Chen 2019; Ng et al. 2019; Lachapelle et al. 2019; Ng, Ghassami, and Zhang 2020; Yu et al. 2021; Gao, Shen, and Xia 2021; Ng et al. 2022). Consequently, differentiable score-based causal discovery methods have garnered significant attention in recent research.

Although current differentiable score-based causal discovery methods have achieved significant success, there are still challenges to be addressed:

1. Current score-based methods use an average scoring approach, assuming that the model's fitting ability is the same for all samples. This assumption neglects the varying difficulty levels among samples, leading to the learning of spurious edges between variables when using average scoring methods (Zhang et al. 2023).
2. Most current score-based methods assume that the collected data is homogeneous, which contradicts real-world situations. When heterogeneous data is present, methods based on this assumption experience a decline in performance (Huang et al. 2020).

Although (Zhang et al. 2023) proposed a model-agnostic adaptive sample reweighting method, it overlooks the presence of outliers in the samples, which limits its effectiveness

*Hongbin Wang is the corresponding author.

in enhancing the performance of current causal discovery methods. Outliers often have disproportionately large loss values. Based on this observation, we believe that samples with loss values in the middle of the distribution should be assigned relatively larger weights.

Inspired by the rapid decay of the Gaussian distribution as distance from the center increases, we propose that a Gaussian function can effectively suppress samples at the two ends of the loss distribution. Additionally, we have enhanced the standard Gaussian with an asymmetric exponential function, allowing the new method to not only suppress extreme samples effectively but also assign relatively larger weights to those in the middle of the distribution. This adaptive weighting method can improve the performance of causal discovery, reduce the impact of spurious edges, and be extended to handle heterogeneous data.

Therefore, we propose a more advanced model-agnostic framework: An asymmetric exponential modulation Gaussian-based adaptive sample reweighting framework, termed DAG-AEG. This framework boosts DAG structure learning by leveraging the distribution of sample losses and using the asymmetric exponential modulation Gaussian to adaptively focus on sample importance. DAG-AEG assesses the importance the importance of each sample based on the errors of the DAG learner, thereby guiding the learner to perform better on more informative samples.

In summary, our contributions are highlighted as:

1. Inspired by the rapid decay of the Gaussian distribution as distance from the center increases, and considering the relationship between sample importance and loss distribution, we propose an improved scheme called AEG (Asymmetric Exponential Modulation Gaussian). This method utilizes an asymmetric exponent to modulate the standard Gaussian, effectively suppressing samples at the extremes of the distribution while assigning relatively greater weights to central samples.
2. We propose a more powerful model-independent framework, called DAG-AEG, which employs Gaussian sample adaptive weighting with asymmetric exponential modulation. By learning the distribution of sample losses, DAG-AEG adaptively emphasizes the importance of samples through asymmetric exponential modulation, thereby enhancing DAG structure learning.
3. Experimental results demonstrate that our proposed DAG-AEG framework significantly boosts causal discovery across both synthetic and real datasets, outperforming existing methods.

Related Works

Causal structure learning is an indispensable and intricate task pervading in various scientifc felds(Liu et al. 2023), which has garnered extensive research attention in recent years. Currently, methods for learning DAG structures are primarily categorized into two types: constraint-based and score-based. Our DAG-AEG primarily focuses on the latter category.

Constraint-based methods initially perform conditional independence tests to obtain the causal skeleton. They then determine the orientations of the edges, refining them up to the Markov equivalence class through established rules. Examples of such methods include the PC algorithm (Spirtes, Glymour, and Scheines 2001), the FCI algorithm (Spirtes, Meek, and Richardson 2013), and the use of kernel-based conditional independence criteria (Zhang et al. 2012). However, these methods lack robustness, as minor errors in constructing the graph skeleton can lead to significant inaccuracies in the inferred Markov equivalence class.

Score-based solutions utilize a scoring function and search strategies to identify the graph that yields the highest score (Ramsey et al. 2017). This approach helps mitigate the shortcomings of constraint-based methods. However, it still faces a significant challenge: the intractable combinatorial nature of the acyclic graph space (Chickering, Heckerman, and Meek 2004). Recently, a groundbreaking study addressed this issue: NOTEARS (Zheng et al. 2018) recasts the combinatorial graph search problem as a continuous optimization problem, resulting in a differentiable score-based optimization method that allows the score function to be optimized through gradient descent. This approach has inspired a substantial body of related literature, such as (Yu et al. 2019; Lachapelle et al. 2019; Zhu, Ng, and Chen 2019; Zheng et al. 2020; Yang et al. 2021; Liu et al. 2023)

Although score-based methods have achieved notable results, they suffer from a significant problem: current methods overly rely on easily fitting samples, which results in spurious edges in the learned causal models. ReScore (Zhang et al. 2023) addresses this by assigning weights to samples based on their corresponding loss—the greater the loss, the more informative the sample, and the higher the assigned weight should be. However, this approach overlooks the presence of outliers in the samples. While this framework can enhance the capability of causal discovery methods to some extent, it still has its flaws.

Preliminaries

Causal Structure Learning

Causal structure learning aims to infer the Structural Equation Model(SEM) from the observational data, which models the data generating procedure (Pearl, Glymour, and Jewell 2016). Formally, let $\mathbf{X} \in R^{n \times d}$ be a sample consisting of n independent and identically distributed observational data of d variables. And we write directed acyclic graphs as $\mathcal{G} = (\mathrm{V}, \mathrm{E})$. Where the nodes V represents the observed variables, denoted as $\mathrm{X} = (X_1, X_2, \cdots X_d)$ and each $\mathrm{edge(i, j)} \in \mathrm{E}$ represents a direct causal relation from X_i to X_j. Given $\mathbf{X}$, we try to learn a DAG $\mathcal{G}$ from a given distribution $\mathrm{P}(\mathrm{X})$. To model X, we can use SEM to model the causal relations between a variable $X_i \in X$ and its parents. This can be formally expressed by Equation 1:

$$X_i = f_i\big(X_{pa(i)}, Z_i\big), \mathrm{i} = 1, 2, \cdots, \mathrm{d} \tag{1}$$

Where $X_{pa(i)}$ denote the parents of X_i, f_i is the causal structure function which can be any linear or nonlinear function, and $Z_i \in Z$ is jointly independent noise variable.

Score-based Causal Structure Learning

Score-based methods assign a score $\mathbf{S}$ to each candidate graph and then search for the best score over the space of all DAGs. This can be formulated as the following combinatorial optimization problem:

$$\min_{\mathcal{G}} S(A) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(x_i, f(A, x_i, \theta)) + \lambda|A|_1$$
$$s.t. \mathcal{G}(A) \in DAGs \tag{2}$$

Where $\mathbf{S}$ is a score function, $\mathcal{G}$ refers to a directed graph, and A is the adjacency matrix of $\mathcal{G}$. While there have been well-defined score function such as the Bayesian Information Criterion (Maxwell Chickering and Heckerman 1997) or Minimum Description Length score (Bouckaert 1993) and the Bayesian Gaussian equivalent score (Geiger and Heckerman 1994), penalized least-squares loss (Zheng et al. 2020), Evidence Lower Bound (Yu et al. 2019), loglikelihood with complexity regularizers (Ng, Ghassami, and Zhang 2020), Maximum Mean Discrepancy (Goudet et al. 2018). Equation 2 is generally NP-hard to solve(Chickering, Heckerman, and Meek 2004), largely due to the combinatorial nature of its acyclicity constraint with the number of DAGs increasing super-exponentially in the number of graph nodes(Zhu, Ng, and Chen 2019). Fortunately, (Zheng et al. 2018) introduced a smooth characterization for the acyclicity converts the combinatorial optimization problem into a continuous constrained optimization problem:

$$\min_{\mathcal{G}} S(A) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(x_i, f(A, x_i, \theta)) + \lambda|A|_1$$
$$s.t. s.t. h(A) = 0 \tag{3}$$

Where $h(A) = tr(e^{A\circ A}) - d = 0$ is a continuous acyclicity constraint. Therefore, this new form of method is referred to as differentiable score-based causal structure learning. To facilitate learning causal graphs, numerous forms(e.g., augmented Lagrangian method) can be applied o solve the Equation. Then, the Equation can be further reformulated as:

$$\min_{\mathcal{G}} S(A) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(x_i, f(A, x_i, \theta)) + \lambda|A|_1 +$$
$$\alpha h(A) + \frac{1}{2}\rho|h(A)|^2 \tag{4}$$

Where α and ρ are coefficients in the Lagrangian method. Although differentiable-based causal discovery methods have achieved notable results, they typically employ an average scoring function (Equation 3). This approach overlooks variations in sample difficulty, which can lead DAG learners to incorrectly infer false edges.

The Proposed DAG-AEG Methodology

To address the issues present in current differentiable score-based methods, we propose a general adaptive sample reweighting framework based on asymmetric exponential modulated Gaussian which boosted DAG structure learning by adaptively focusing on the importance of samples through modulated Gaussian. For clarity, we first present the principles of this framework. Afterward, we describe the approach of applying it to DAG structure learning.

The Theory of Asymmetric Exponential Modulation Gaussian

In current differentiable score-based causal discovery approaches, applying the average score function uniformly across all samples can lead to the DAG learner overfitting those samples that are easier to fit, thereby introducing false edges. ReScore(Zhang et al. 2023) proposes that even though the importance of each sample is unknown, weights can be assigned based on the relative ease or difficulty with which the DAG learner fits the samples:

$$\min_{\mathcal{G}} S_w(A) = \sum_{i=1}^{n} w_i\mathcal{L}(x_i, f(A, x_i, \theta)) + \lambda|A|_1$$
$$+ \alpha h(A) + \frac{1}{2}\rho|h(A)|^2 \tag{5}$$

Where $\mathrm{w} = (w_1, w_2, ..., w_n)$ is a sample reweighting vector. In ReScore, the authors assign larger weights w_i to samples that are more challenging to fit. While this enhances the performance of the DAG learner to some extent, it overlooks the presence of outliers in the samples, still posing certain limitations.

There is a general consensus that neural network models tend to focus on simpler samples that are easier to fit, meaning that samples with smaller loss values often contribute less to the overall model training. Conversely, samples with larger losses, which are deemed more challenging, tend to contribute more significantly (Li, Liu, and Wang 2019). This principle is similarly applicable in the learning of DAG structures. However, the presence of outliers in the data complicates this scenario. Outliers typically exhibit higher loss values and are often found at the extremes of the sample loss distribution. Consequently, a high loss value does not necessarily warrant a higher weight in the learning process. In light of this, it is necessary to moderate the influence of samples at both extremes of the loss distribution—those with very small or very large losses—while assigning relatively greater weights to samples that are distributed more centrally within the loss distribution.

Inspired by the rapid decay of a Gaussian distribution as the distance from the center increases, we believe that the Gaussian function can effectively reduce the influence of extreme values at both ends of the sample distribution, which can be expressed in this form:

$$\mathrm{w} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\mathcal{L}^2 - \mu}{\sigma}\right)^2\right) \tag{6}$$

where $\mathcal{L}$ represents the reconstruction loss of sample $\mathbf{x}$, μ represents the intermediate value of the batch sample loss, and σ is the hyperparameter.

However, the standard Gaussian distribution is symmetric around its central point. To follow the consensus that samples with higher loss values within a specific range are often more important in the sample weighting process, we need to

introduce a bias function. Here, we use an asymmetric exponential function as the bias for the standard Gaussian. This new approach effectively suppresses samples at both ends while giving relatively greater weights to the middle samples. We refer to this modified Gaussian distribution as The Asymmetric Exponential Modulation Gaussian (AEG):

$$\mathrm{w} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathcal{L}^2-\mu}{\sigma}\right)^2\right) + \beta(\mathcal{L}^2-\mu)\exp\left(-\gamma(\mathcal{L}^2-\mu)\right) * 1_{\{\mathcal{L}^2>\mu\}} \quad (7)$$

The indicator function $1_{\{\mathcal{L}^2>\mu\}}$ is effective only when the sample loss is above the median. The parameters β and γ are dynamically adjusted based on the loss magnitude. The core improvement lies in minimizing the contribution of samples with extreme loss values—either very high or very low (easily fitted samples and outliers)—to the DAG learner by assigning them smaller weights. Conversely, samples with intermediate loss values contribute more to the DAG learner. The larger their loss, the more they contribute, providing additional insight into depicting causal edges, thus necessitating larger weights for these samples.

Boosting Causal Structure Learning via AEG-Based Adaptive Sample Reweighting

We believe that samples with either excessively high or excessively low losses are not ideal for guiding the DAG learner. The primary reason is that samples with a high degree of fit offer limited critical information, while samples with a low degree of fit are more likely to be outliers that the model cannot accurately fit. In contrast, samples with losses that fall in the middle range typically contain a substantial amount of critical information. In the previous section, we discussed the weight adaptation method. In this section, we will apply this method to causal structure learning to enhance its performance.

Therefore, we propose a more powerful model-agnostic framework, called DAG-AEG, an adaptive sample reweighting framework based on AEG. This framework boosts DAG structure learning by analyzing the distribution of sample losses and employing the proposed method for adaptive sample attention. We adopt a Bi-Level learning approach, where the learned weights w are used to recalculate scores. Specifically, this framework is applied to causal structure learning, represented in the following form:

$$\begin{aligned}
\min_{\mathcal{G}} S_w(A) &= \sum_{i=1}^{n} w_i \mathcal{L}(x_i, f(A, x_i, \theta_{\mathcal{G}})) + \lambda|A|_1 \\
&\quad + \alpha h(A) + \frac{1}{2}\rho|h(A)|^2, \\
&\text{s.t.} w \in argmax \sum_{i=1}^{n} w_i \mathcal{L}(x_i, f(A, x_i, \theta_{\mathcal{G}})) \\
w_{i|L_i\in\mathcal{L}} &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{L_i^2-\mu}{\sigma}\right)^2\right) + \\
&\quad \beta({L_i}^2-\mu)\exp\left(-\gamma({L_i}^2-\mu)\right) * 1_{\{{L_i}^2>\mu\}}
\end{aligned} \quad (8)$$

Algorithm 1: The proposed DAG-AEG framework to differentiable score-based causal discovery

Input: Observed data $\mathbf{X} \in R^{n\times d}$, reweighting model parameters θ_w, maximum epoch in the inner loop K_{inner}, maximum epoch in the outer loop K_{outer}
Initialize: initialize θ_w to uniformly output $\frac{1}{n}$
Output: predicted $\mathcal{G}$

for $k_1 = 0$ to K_{outer} **do**
  Fix reweighting model parameters θ_w;
  Get w through reweighting model utilizing the sample loss values calculated by the DAG learner;
  Optimize DAG learner parameters $\theta_{\mathcal{G}}$ by minimizing $\min_{\mathcal{G}} S_w(A)$;
  if $k_1 \geq start \quad reweighting \quad epoch$ **then**
    for $k_2 = 0$ to K_{inner} **do**
      Fix DAG learner parameters $\theta_{\mathcal{G}}$;
      Calculate w through reweighting model in equation 7;
      Optimize θ_w by maximizing $\sum_{i=1}^{n} w_i \mathcal{L}(x_i, f(A, x_i, \theta_{\mathcal{G}}))$;
      $k_2 \leftarrow k_2 + 1$
    end for
    $k_1 \leftarrow k_1 + 1$
    $k_2 \leftarrow 0$
  end if
end for
return predicted $\mathcal{G}$

The formula includes two objectives, with the lower-level objective nested within the upper-level objective. In the lower-level loop, the DAG learner is fixed. Based on the sample loss distribution learned by the DAG learner, the reweighted scoring function is learned using the asymmetric exponential modulation Gaussian. In the upper-level loop, this reweighted scoring function is minimized to optimize the DAG learner on the reweighted observational data determined in the lower-level loop. By alternately training these upper and lower loops, the importance of each sample is adaptively assessed according to the error of the DAG learner. This process gradually guides the DAG learner to perform better on informative samples and to learn an optimal causal structure model, as detailed in the algorithm 1. As a general sample adaptive weighting framework, DAG-AEG can be applied to any differentiable score-based causal structure learning methods.

Experiments

In this section, we conduct experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our method DAG-AEG. We are interested in how well it can 1)broadly enhance the differentiable score-based causal discovery baselines and 2)perform when dealing with heterogeneous data.

Experimental Settings

Baselines. **To evaluate its performance in enhancing existing baselines,** we selected four state-of-the-art causal dis-

d	METHODS	ER2				ER4			
		TPR↑	FDR↓	SHD↓	SID↓	TPR↑	FDR↓	SHD↓	SID↓
10	NOTEARS	0.85±0.09	0.07±0.07	5.8±2.2	20.8±5.2	0.79±0.11	0.09±0.05	10.0±5.2	25.8±9.9
	+ ReScore	0.89±0.07$^{+5\%}$	0.08±0.09$^{-14\%}$	4.6±2.3$^{+21\%}$	12.8±7.0$^{+39\%}$	0.85±0.04$^{+8\%}$	**0.05**±0.04$^{+44\%}$	7.2±1.9$^{+28\%}$	24.2±8.4$^{+6\%}$
	+ DAG-AEG	0.94±0.06$^{+11\%}$	**0.05**±0.06$^{+29\%}$	**2.0**±1.9$^{+66\%}$	**6.6**±8.4$^{+68\%}$	**0.89**±0.05$^{+13\%}$	0.06±0.06$^{+33\%}$	**5.7**±2.9$^{+43\%}$	**6.6**±8.4$^{+74\%}$
	GOLEM	0.87±0.06	0.22±0.11	6.5±3.4	13.0±6.7	0.63±0.03	0.16±0.03	17.2±1.3	48.0±13.3
	+ ReScore	0.88±0.06$^{+1\%}$	0.21±0.11$^{+5\%}$	6.0±3.4$^{+8\%}$	12.4±6.3$^{+5\%}$	0.66±0.06$^{+5\%}$	0.17±0.01$^{-6\%}$	16.2±1.0$^{+6\%}$	46.7±13.3$^{+3\%}$
	+ DAG-AEG	**0.95**±0.04$^{+9\%}$	0.15±0.11$^{+32\%}$	3.6±2.5$^{+45\%}$	7.8±8.1$^{+40\%}$	0.82±0.08$^{+30\%}$	0.08±0.03$^{+50\%}$	9.4±3.2$^{+45\%}$	26.3±8.8$^{+45\%}$
20	NOTEARS	0.85±0.08	**0.09**±0.03	9.2±3.8	55.4±31.1	0.74±0.02	**0.23**±0.03	39.4±7.9	185.8±38.1
	+ ReScore	0.87±0.07$^{+2\%}$	0.11±0.05$^{-22\%}$	8.8±3.5$^{+4\%}$	50.6±26.3$^{+9\%}$	0.79±0.05$^{+7\%}$	0.28±0.05$^{-22\%}$	36.8±7.9$^{+7\%}$	122.7±40.1$^{+34\%}$
	+ DAG-AEG	0.91±0.02$^{+7\%}$	0.10±0.03$^{-11\%}$	**7.2**±1.5$^{+22\%}$	35.2±14.3$^{+37\%}$	**0.86**±0.02$^{+16\%}$	0.23±0.05$^{+0\%}$	**30.6**±6.1$^{+22\%}$	**122.7**±40.1$^{+34\%}$
	GOLEM	0.75±0.12	0.20±0.11	17.0±6.1	78.2±22.6	0.46±0.06	0.50±0.06	73.6±7.9	249.8±7.8
	+ ReScore	0.76±0.06$^{+1\%}$	0.20±0.10$^{+0\%}$	15.8±5.8$^{+7\%}$	77.0±21.5$^{+2\%}$	0.48±0.06$^{+4\%}$	0.43±0.10$^{+14\%}$	70.2±5.8$^{+5\%}$	246.2±11.4$^{+1\%}$
	+ DAG-AEG	**0.96**±0.03$^{+28\%}$	0.20±0.10$^{+0\%}$	11.3±8.8$^{+34\%}$	**17.6**±18.2$^{+78\%}$	0.57±0.20$^{+24\%}$	0.39±0.04$^{+22\%}$	61.1±8.7$^{+17\%}$	187.0±66.2$^{+25\%}$
50	NOTEARS	0.79±0.06	**0.09**±0.03	27.6±7.7	427.0±186.1	0.51±0.12	0.27±0.10	133.4±29.5	1644±172
	+ ReScore	0.88±0.06$^{+11\%}$	0.15±0.04$^{-67\%}$	26.2±7.6$^{+5\%}$	266.0±146.4$^{+38\%}$	0.52±0.21$^{+2\%}$	0.29±0.07$^{-7\%}$	130.2±37.4$^{+2\%}$	1454±337$^{+12\%}$
	+ DAG-AEG	0.88±0.05$^{+11\%}$	0.14±0.04$^{-56\%}$	**25.0**±6.2$^{+9\%}$	255.3±117.5$^{+40\%}$	**0.69**±0.13$^{+35\%}$	**0.25**±0.07$^{+7\%}$	**104.8**±29.7$^{+21\%}$	**1191**±282$^{+28\%}$
	GOLEM	0.80±0.09	0.35±0.09	68.6±19.7	433.5±215.6	0.31±0.11	0.68±0.06	150.6±25.1	1775±162
	+ ReScore	0.82±0.15$^{+3\%}$	0.33±0.14$^{+6\%}$	63.4±27.9$^{+8\%}$	430.2±155.5$^{+1\%}$	0.39±0.06$^{+26\%}$	0.66±0.06$^{+3\%}$	146.3±26.3$^{+3\%}$	1644±115$^{+7\%}$
	+ DAG-AEG	**0.91**±0.14$^{+14\%}$	0.32±0.16$^{+9\%}$	51.7±29.4$^{+25\%}$	**173.2**±258.0$^{+60\%}$	0.40±0.16$^{+29\%}$	0.56±0.14$^{+18\%}$	136.1±32.1$^{+10\%}$	1496±131$^{+16\%}$

Table 1: Results of linear models for ER graphs with 10,20 and 50 nodes.

covery methods: two for linear systems (NOTEARS(Zheng et al. 2018), GOLEM(Ng, Ghassami, and Zhang 2020)) and two for nonlinear settings (NOTEARS-MLP(Zheng et al. 2020), GraN-DAG(Lachapelle et al. 2019)). We integrated our DAG-AEG framework with these methods and compared it against the four baselines. **In terms of handling heterogeneous data,** we compared GOLEM+DAG-AEG and NOTEARS-MLP+DAG-AEG to the state-of-the-art baseline CD-NOD (Huang et al. 2020) and the recently proposed approach DICD(Wang et al. 2022). In both cases, we also compared DAG-AEG with the current model-agnostic state-of-the-art method ReScore(Zhang et al. 2023). The detailed configuration of hyperparameters for each model is comprehensively documented in the "Hyperparameter Design" section of the technical appendix.

Datasets. To validate the performance of our framework on the aforementioned problems 1) and 2), we follow the convention of causal discovery and use the same experimental setup as in (Zhang et al. 2023), specifically as follows:

- **To assess the performance in enhancing existing baselines,** we employ a well-known graph sampling model: Erdos-Renyi (ER) to generate random DAGs. We varied the number of variables($d = \{10, 20, 50\}$) with edge density ($degree = \{2, 4\}$,denoted as ERk or SFk). For each graph, we generate 10 datasets of 2,000 samples. For the linear settings, similar to (Zheng et al. 2018) and(Gao, Shen, and Xia 2021), the coefficients are assigned following Uniform distribution $U(-2, -0.5) \cup U(0.5, 2)$ with additive standard Gaussian noise. For the nonlinear settings, as in (Zheng et al. 2020), we generate the ground truth SEM using Equation 1, under the Gaussian process (GP) with a radial basis function kernel of bandwidth one, where $f_i(\cdot)$ is additive noise model with Z_i as an i.i.d. random variable following a standard normal distribution. Notice that both of these settings are known to be fully identifiable (Peña 2018; Peters et al. 2014). In addition, we report the mean and standard deviations of the metrics to ensure a fair comparison.
- **To evaluate performance on heterogeneous data,** we used both synthetic and real-world heterogeneous data.
  - **Synthetic heterogeneous data:** We considered both linear and nonlinear settings (n = 1000, d = 20, ER2) containing two distinct groups. 10% of observations come from a disadvantaged group, where half of the noise variables Z_i (defined in Equation 1) follow $N(0, 1)$ and the other half follow $N(0, 0.1)$. Conversely, 90% of observations are from a dominant group where the scales of noise variables are reversed.
  - **Real-world heterogeneous data:** We used the well-known Sachs dataset (Sachs et al. 2005), which measures the levels of various proteins and phospholipids in human cells under nine different perturbation conditions, each involving specific reagents. With perturbation conditions annotated, we treated Sachs as real-world heterogeneous data(Mooij, Magliacane, and Claassen 2020; Zhang et al. 2023). The ground truth causal graph for this dataset includes 11 variables and 17 edges. Our tests were conducted on observational data comprising 7466 samples.

Evaluation Metrics. To evaluate the performance of DAG structure learning, we consider four metrics: True Positive Rate (TPR), False Discovery Rate (FDR), Structural Hamming Distance (SHD), and Structural Intervention Distance (SID) (Peters and Bühlmann 2015), averaged over ten random trials. The SHD represents the minimum number of edge additions, deletions, and reversals needed to convert the estimated graph into the true DAG, encompassing both false positives and false negatives. The SID measures the number of interventional distributions that differ between the true and recovered networks. For optimal performance,

d	METHODS	ER2				ER4			
		TPR↑	FDR↓	SHD↓	SID↓	TPR↑	FDR↓	SHD↓	SID↓
10	NOTEARS-MLP	0.76±0.17	0.14±0.09	7.0±3.5	17.9±10.0	0.83±0.05	0.21±0.04	10.9±1.9	28.6±12.0
	+ ReScore	0.73±0.07$^{-4\%}$	0.10±0.09$^{+29\%}$	6.8±2.9$^{+3\%}$	20.3±9.7$^{-13\%}$	0.94±0.06$^{+13\%}$	0.15±0.06$^{+29\%}$	6.8±2.7$^{+38\%}$	8.8±12.4$^{+69\%}$
	+ DAG-AEG	0.77±0.11$^{+1\%}$	0.09±0.06$^{+36\%}$	5.9±1.9$^{+16\%}$	16.5±7.4$^{+8\%}$	0.98±0.02$^{+18\%}$	0.11±0.02$^{+48\%}$	5.2±1.1$^{+52\%}$	3.6±5.3$^{+87\%}$
	GraN-DaG	0.88±0.06	0.02±0.03	2.7±1.6	8.7±4.8	0.98±0.02	0.12±0.03	5.4±1.1	3.7±4.8
	+ ReScore	0.90±0.05$^{+2\%}$	0.01±0.03$^{+50\%}$	2.4±1.1$^{+11\%}$	7.2±3.0$^{+17\%}$	**0.99**±0.01$^{+1\%}$	0.11±0.01$^{+8\%}$	**4.8**±0.6$^{+11\%}$	**0.5**±0.81$^{+87\%}$
	+ DAG-AEG	**0.90**±0.04$^{+2\%}$	**0.01**±0.02$^{+50\%}$	**2.2**±0.9$^{+19\%}$	**6.5**±3.6$^{+25\%}$	0.99±0.02$^{+1\%}$	**0.10**±0.01$^{+17\%}$	4.9±0.3$^{+9\%}$	0.7±1.3$^{+81\%}$
20	NOTEARS-MLP	0.70±0.12	0.13±0.07	14.9±5.4	98.4±22.5	0.44±0.09	0.26±0.10	55.0±9.2	176.3±33.3
	+ ReScore	0.73±0.09$^{+4\%}$	0.11±0.05$^{+15\%}$	13.7±5.1$^{+8\%}$	88.8±23.8$^{+10\%}$	0.41±0.07$^{-7\%}$	0.17±0.08$^{+35\%}$	51.6±6.4$^{+6\%}$	179.9±33.7$^{-2\%}$
	+ DAG-AEG	0.74±0.09$^{+6\%}$	0.12±0.07$^{+8\%}$	13.4±5.2$^{+10\%}$	77.6±39.0$^{+32\%}$	0.45±0.05$^{+2\%}$	0.20±0.05$^{+23\%}$	51.9±4.1$^{+6\%}$	164.2±23.8$^{+7\%}$
	GraN-DaG	0.81±0.15	0.08±0.08	9.3±5.4	53.4±24.4	0.20±0.07	0.18±0.08	57.4±4.6	131.5±21.4
	+ ReScore	0.81±0.14$^{+0\%}$	0.05±0.04$^{+38\%}$	8.5±5.7$^{+9\%}$	51.0±24.6$^{+5\%}$	0.21±0.07$^{+5\%}$	0.17±0.09$^{+6\%}$	56.2±4.6$^{+2\%}$	**125.4**±23.3$^{+5\%}$
	+ DAG-AEG	**0.83**±0.09$^{+3\%}$	**0.01**±0.02$^{+88\%}$	**7.4**±3.3$^{+20\%}$	**42.3**±11.5$^{+21\%}$	**0.53**±0.06$^{+165\%}$	**0.11**±0.06$^{+39\%}$	**42.9**±4.7$^{+25\%}$	136.5±30.0$^{-4\%}$
50	NOTEARS-MLP	0.32±0.04	0.13±0.08	69.5±4.7	884.4±172.8	0.17±0.02	0.06±0.04	167.0±4.1	1608±97
	+ ReScore	0.51±0.08$^{+59\%}$	0.10±0.07$^{+23\%}$	53.5±8.7$^{+23\%}$	628.1±120.6$^{+29\%}$	0.26±0.04$^{+53\%}$	0.11±0.05$^{-83\%}$	154.4±6.4$^{+8\%}$	1438±111$^{+11\%}$
	+ DAG-AEG	**0.62**±0.10$^{+94\%}$	0.09±0.02$^{+31\%}$	42.1±9.1$^{+39\%}$	**505.8**±123.4$^{+43\%}$	0.27±0.05$^{+59\%}$	0.06±0.03$^{+0\%}$	149.4±9.6$^{+11\%}$	1400±138$^{+13\%}$
	GraN-DaG	0.52±0.09	0.15±0.05	51.6±9.3	632.8±140.3	0.32±0.04	0.08±0.16	141.6±8.2	1379±91
	+ ReScore	0.53±0.06$^{+2\%}$	0.11±0.02$^{+28\%}$	46.0±6.0$^{+11\%}$	581.0±104.7$^{+8\%}$	0.31±0.03$^{-3\%}$	0.06±0.04$^{+25\%}$	138.8±7.5$^{+2\%}$	1351±98$^{+2\%}$
	+ DAG-AEG	0.61±0.03$^{+17\%}$	**0.06**±0.03$^{+60\%}$	**40.3**±3.1$^{+22\%}$	572.5±107.9$^{+10\%}$	**0.33**±0.03$^{+3\%}$	**0.05**±0.02$^{+38\%}$	**136.0**±9.6$^{+4\%}$	**1249**±114$^{+9\%}$

Table 2: Results of nonlinear models for ER graphs with 10,20 and 50 nodes.

TPR should be high, while FDR, SHD, and SID should be low, indicating a more accurate estimate of the target causal graph.

Result Analysis

Performance on Baseline Enhancement. In this section, we present the experimental results of DAG-AEG and compare them with the baselines and the model-independent ReScore on the previously introduced synthetic datasets in terms of TPR, FDR, SHD, and SID metrics. Tables 1 and 2 show the empirical results for both linear and nonlinear synthetic data. The error bars represent the standard deviation across datasets over ten trials. Red percentages indicate performance increases, while blue percentages indicate decreases, of the model-agnostic frameworks compared to the original score-based methods for each metric. The best-performing methods are highlighted in bold. Our findings include:

DAG-AEG consistently enhances score-based DAG structure learning methods across all datasets, outperforming the model-agnostic state-of-the-art method, ReScore. Unlike ReScore, DAG-AEG shows rarely performance degradation in terms of SHD and SID. Specifically, it improves SHD by approximately 3% to 66% over advanced baselines, and by 0 to 40% over ReScore, with fewer missing, erroneously detected, and reversed edges. For SID, although there is one case of decline, overall it achieves improvements of 7% to 87% over advanced baselines, surpassing ReScore. We attribute these improvements to AEG, which suppresses the tails of the loss distribution while identifying significant samples, thus improving the quality of score-based DAG learners. Furthermore, a detailed examination of TPR and FDR reveals that DAG-AEG lowers FDR by eliminating spurious edges and increases TPR by identifying more correct edges, outperforming ReScore. This indicates that DAG-AEG more effectively filters and increases the weight of information-rich samples while suppressing the weight of less informative ones, thus better extracting causal relationships.

TYPE	METHODS	TPR↑	FDR↓	SHD↓
linear	GOLEM	0.79	0.33	18.7
	+ IPS	0.65	0.19	18.6
	+ ReScore	0.81	0.24	16.4
	+ DAG-AEG	**0.93**	0.23	**15.1**
	CD-NOD	0.51	**0.17**	24.1
	DICD	0.82	0.28	16.7
nonlinear	NOTEARS-MLP	0.62	0.36	25.8
	+ IPS	0.35	**0.21**	28.7
	+ ReScore	0.63	0.32	23.8
	+ DAG-AEG	**0.65**	0.30	**22.0**
	CD-NOD	0.60	0.29	26.0
	DICD	0.50	0.24	23.5

Table 3: Results on synthetic heterogeneous data.

Differentiable score-based baseline methods show significant performance degradation on dense graphs. As illustrated in the tables, their performance diminishes as the number of nodes and edges in the DAG increases. This decline is particularly evident with 50 vertices and 100 to 200 edges, where the baseline methods perform poorly. The issue worsens with non-linear models, as the TPR drops below 50%. This can be clearly seen in Figure 1. This degradation is mainly due to the increasing difficulty of enforcing the acyclic constraint as graph density increases (Charpentier, Kibler, and Günnemann 2022; Chen, Wu, and Jin 2024). Figure 1 illustrates that, with a constant degree of 4, the performance of the baseline combined with either ReScore or DAG-AEG declines sharply as the number of nodes increases. Despite this decline, our proposed method consistently provides the best improvement to the baseline. This

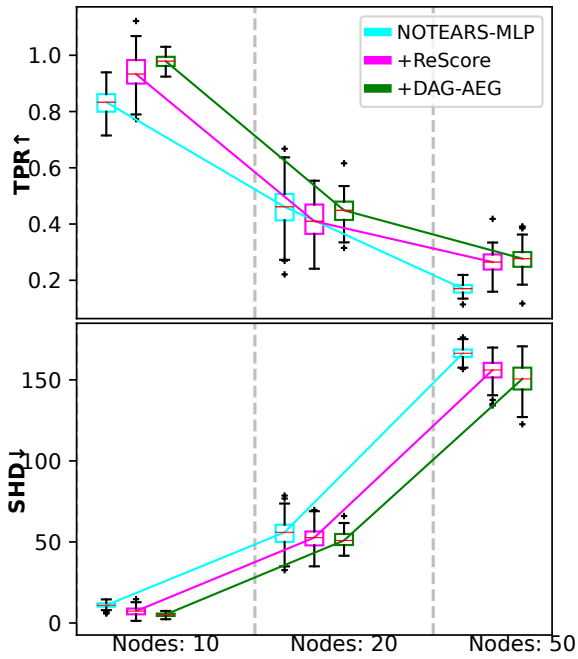

Figure 1: TPR and SHD of NOTEARS, the combination of ReScore, and the combination of DAG-AEG on graphs generated using the ER model with a degree of 4 and vertex counts of 10, 20, and 50, respectively.

trend is further supported by the comprehensive metrics presented in Table 1 and Table 2. Overall, although the DAG-AEG framework can improve the performance of baseline methods, its effectiveness still depends on the efficiency of the DAG learner.

Performance on Heterogeneous Data. In this part, we compare Baseline+DAG-AEG with two group annotation-dependent causal learning methods to validate DAG-AEG's capability in handling heterogeneous data. Additionally, we consider a Baseline+IPS reweighting method, where sample weights are inversely proportional to group sizes. We also compare the performance of DAG-AEG with ReScore. The experiments are conducted on both synthetic and real heterogeneous data, with the specific analysis detailed below:

Performance on synthetic heterogeneous data: In Table 3, our DAG-AEG framework significantly improves the TPR, FDR, and SHD metrics of baseline models on synthetic heterogeneous data, irrespective of using linear or non-linear models. Notably, the TPR and SHD metrics achieve optimal values for their respective types (bolded in the table). It is important to note that the baseline models used here are not specifically designed for heterogeneous data. Additionally, DAG-AEG enhanced baselines outperform recognized CD-NOD and DICD lower bounds in the SHD metric. Compared to the fixed-weight IPS method, the IPS+baseline method shows a severe drop in TPR performance, highlighting the importance of adaptive weights. Among adaptive weighting methods, although ReScore enhances baseline performance, it still falls short compared to our DAG-AEG framework, underscoring the competitiveness of the proposed AEG.

METHODS	TPR↑	FDR↓	SHD↓	SID↓	#PE
GOLEM	0.176	**0.026**	15	53	22
+ ReScore	0.294	0.063	14	49	6
+ DAG-AEG	**0.294**	0.063	**13**	**47**	6
NPTEARS-MLP	0.412	0.632	16	45	19
+ ReScore	0.412	0.500	13	43	14
+ DAG-AEG	**0.412**	**0.462**	**13**	**42**	14
GraN-DAG	0.294	0.643	16	60	14
+ ReScore	0.353	0.600	15	58	15
+ DAG-AEG	**0.471**	**0.556**	**14**	**38**	18
GES	0.294	0.853	31	54	34
+ ReScore	0.588	0.722	28	50	36
+ DAG-AEG	**0.647**	**0.667**	**24**	**26**	33
CD-NOD	0.588	0.444	15	-	18

Table 4: Results on Sachs dataset.

Performance on real heterogeneous data: As shown in Table 4, DAG-AEG significantly improves the TPR, FDR, SHD, and SID metrics across all baseline methods on real heterogeneous datasets. The enhanced TPR and FDR indicate that DAG-AEG predicts more correct edges and fewer incorrect edges. Compared to CD-NOD, which uses annotations as prior knowledge for heterogeneous data, our DAG-AEG+GraN-DAG achieves a competitive TPR without requiring ground-truth annotations. Moreover, DAG-AEG+GraN-DAG outperforms CD-NOD in SHD when predicting the same number of edges(#PE). Compared to ReScore, DAG-AEG shows superior performance. Experiments on both real and synthetic heterogeneous data demonstrate the practical competitiveness of our proposed DAG-AEG framework.

Conclusions

Learning causal structures from observational data poses significant challenges. Despite the good results achieved by current differentiable scoring-based methods, the causal structures they identify often deviate from actual conditions. In this paper, we introduce DAG-AEG, an innovative model-agnostic framework that boosts DAG structure learning. Utilizing the distribution of sample losses, DAG-AEG dynamically prioritizes sample importance through the AEG technique. This approach not only dynamically adjusts sample weights to underscore their relevance but also significantly enhances model performance across heterogeneous datasets. Comprehensive experiments demonstrate that DAG-AEG substantially improves the performance of existing scoring-based linear and nonlinear DAG learners across a range of synthetic and real-world datasets, thereby highlighting its practical efficacy and competitive edge. In future research, we aim to enhance the performance of complex dense graphs by refining acyclic constraints and effectively integrating the findings from this study.

Acknowledgments

This work was supported by the Basic Research Project(No.JCKY2022203B001).

References

Boruch, R. F. 1997. *Randomized experiments for planning and evaluation: A practical guide*, volume 44. Sage.

Bouckaert, R. R. 1993. Probabilistic network construction using the minimum description length principle. In *European conference on symbolic and quantitative approaches to reasoning and uncertainty*, 41–48. Springer.

Castro, D. C.; Walker, I.; and Glocker, B. 2020. Causality matters in medical imaging. *Nature Communications*, 11(1): 3673.

Charpentier, B.; Kibler, S.; and Günnemann, S. 2022. Differentiable dag sampling. *arXiv preprint arXiv:2203.08509*.

Chen, S.; Wu, H.; and Jin, G. 2024. Causal structure learning for high-dimensional non-stationary time series. *Knowledge-Based Systems*, 295: 111868.

Chickering, M.; Heckerman, D.; and Meek, C. 2004. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5: 1287–1330.

Gao, Y.; Shen, L.; and Xia, S.-T. 2021. Dag-gan: Causal structure learning with generative adversarial nets. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3320–3324. IEEE.

Geiger, D.; and Heckerman, D. 1994. Learning gaussian networks. In *Uncertainty in Artificial Intelligence*, 235–243. Elsevier.

Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, 39–80.

Huang, B.; Zhang, K.; Zhang, J.; Ramsey, J.; Sanchez-Romero, R.; Glymour, C.; and Schölkopf, B. 2020. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89): 1–53.

Lachapelle, S.; Brouillard, P.; Deleu, T.; and Lacoste-Julien, S. 2019. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*.

Li, B.; Liu, Y.; and Wang, X. 2019. Gradient harmonized single-stage detector. In *Proceedings of the AAAI conference on artificial intelligence*, 01, 8577–8584.

Li, J.; Liu, L.; Le, T. D.; and Liu, J. 2020. Accurate data-driven prediction does not mean high reproducibility. *Nature machine intelligence*, 2(1): 13–15.

Liang, J.; Wang, J.; Yu, G.; Xia, S.; and Wang, G. 2024. Multi-Granularity Causal Structure Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12, 13727–13735.

Liu, F.; Ma, W.; Zhang, A.; Wang, X.; Duan, Y.; and Chua, T.-S. 2023. Discovering Dynamic Causal Space for DAG Structure Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1429–1440.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

Maxwell Chickering, D.; and Heckerman, D. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine learning*, 29: 181–212.

Mooij, J. M.; Magliacane, S.; and Claassen, T. 2020. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99): 1–108.

Ng, I.; Ghassami, A.; and Zhang, K. 2020. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33: 17943–17954.

Ng, I.; Zhu, S.; Chen, Z.; and Fang, Z. 2019. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*.

Ng, I.; Zhu, S.; Fang, Z.; Li, H.; Chen, Z.; and Wang, J. 2022. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 424–432. SIAM.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

Pearl, J. 2019. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3): 54–60.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peña, J. M. 2018. Identifiability of gaussian structural equation models with dependent errors having equal variances. *arXiv preprint arXiv:1806.08156*.

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

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models. *Machine Learning Research*.

Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; and Glymour, C. 2017. 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: 121–129.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529.

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, prediction, and search*. MIT press.

Spirtes, P. L.; Meek, C.; and Richardson, T. S. 2013. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*.

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2022. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4): 1–36.

Wang, Y.; Zhang, A.; Wang, X.; Yuan, Y.; He, X.; and Chua, T.-S. 2022. Differentiable invariant causal discovery. *arXiv preprint arXiv:2205.15638*.

Xu, G.; Duong, T. D.; Li, Q.; Liu, S.; and Wang, X. 2020. Causality learning: A new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789*.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9593–9602.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International conference on machine learning*, 7154–7163. PMLR.

Yu, Y.; Gao, T.; Yin, N.; and Ji, Q. 2021. DAGs with no curl: An efficient DAG structure learning approach. In *International Conference on Machine Learning*, 12156–12166. Pmlr.

Zhang, A.; Liu, F.; Ma, W.; Cai, Z.; Wang, X.; and Chua, T.-S. 2023. Boosting Differentiable Causal Discovery via Adaptive Sample Reweighting. *arXiv preprint arXiv:2303.03187*.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

Zheng, X.; Dan, C.; Aragam, B.; Ravikumar, P.; and Xing, E. 2020. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, 3414–3425. Pmlr.

Zhu, S.; Ng, I.; and Chen, Z. 2019. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*.