Improving Gradient-based DAG Learning by Structural Asymmetry

Yujie Wang
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
yujiewang@mail.hfut.edu.cn

Shuai Yang
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
yangs@mail.hfut.edu.cn

Xianjie Guo
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
xianjieguo@mail.hfut.edu.cn

Kui Yu
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
yukui@hfut.edu.cn

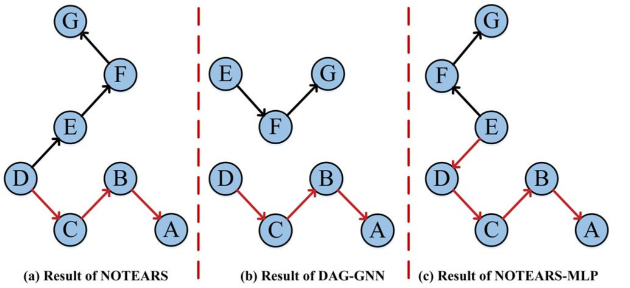
2021 IEEE International Conference on Big Knowledge (ICBK) | 978-1-6654-3858-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICKG52313.2021.00022

***Abstract*—Directed acyclic graph (DAG) learning plays a fundamental role in causal inference and other scientific scenes, which aims to uncover the relationships between variables. However, identifying a DAG from observational data has always been a challenging task. Recently, gradient-based DAG learning algorithms that convert a combination-optimization DAG learning problem into a continuous-optimization problem have achieved emerging successes. These algorithms are easy to optimize and able to deal with both parametric and non-parametric data but suffer from many reversed edges learnt by these algorithms. In this paper, we propose a framework named Residual Independence Test (RIT) to correct those reversed edges by leveraging the structural asymmetry reflected in the dependence between regression residual and direct cause. We conduct extensive experiments on both synthetic and benchmark datasets, the results show that the RIT framework significantly improve the performance of gradient-based DAG learning algorithms.**

***Index Terms*—Directed acyclic graph, Structural asymmetry, Gradient-based structure learning**

I. Introduction

Directed Acyclic Graph (DAG) learning plays an essential role in causal inference [1], [2], machine learning [3]–[5], [6] and explainable model [7], [8]. Observational DAG learning aims to learn DAGs from purely observational data. Traditional DAG learning methods can be divided into constraint-based and score-based categories. Constraint-based methods [9], [10], [11] firstly learn the skeleton of a DAG by performing conditional independence tests and then orient the edges using certain rules. Meanwhile, score-based methods [12], [13] use score functions to evaluate the quality of candidate DAGs and discover the DAG with the highest score. Recently, a new class of DAG learning algorithms based on gradient descent has been proposed, which adopts numerical optimization methods or exploits the powerful modeling ability of neural networks to learn DAGs. The core idea of these approaches is minimizing the least square loss between original samples and reconstructed samples or promoting the value of maximum likelihood estimation.



(a) Result of NOTEARS (b) Result of DAG-GNN (c) Result of NOTEARS-MLP

Fig. 1: (a), (b) and (c) show the results of NOTEARS, DAG-GNN, and NOTEARS-MLP, respectively (black lines denote the correct edges and red lines represent the reversed edges learnt by the three algorithms)

However, the DAG learnt by gradient-based methods is not so accurate as expected. Several methods do not satisfy the important assumptions for identifiability [14], [15], hence they cannot learn an accurate DAG. In other words, these methods can discover potential causal relationships between variables, but cannot effectively distinguish direct causes and direct effects, resulting in that there exist many reversed edges in the learnt DAG. For example, we run three state-of-the-art gradient-based methods, NOTEARS [16], NOREARS-MLP [17], and DAG-GNN [18] on the chain network [19], a benchmark Bayesian network with 7 nodes and 6 edges. We use black lines and red ones to denote the correct edges and the reversed edges learnt by the three algorithms, respectively. From the results shown in Fig. 1, we find that the DAGs learnt by NOTEARS, DAG-GNN and NOTEARS-MLP have three, three, and four reversed edges, respectively. Therefore, the presence of reversed edges is a serious problem in gradient-based methods.

Methods for determining the causal directions between pairwise variables mainly rely on structural asymmetry. Regression

978-1-6654-3858-2/21/$31.00 ©2021 IEEE
DOI 10.1109/ICKG52313.2021.00022

Error based Causal Inference (RECI) [20] fits regression models in both possible directions, and hence the true causal direction has a smaller least-squares error. Information-Geometric Causal Inference (IGCI) [21] uses entropy to discover the asymmetry between the direct cause and the direct effect, further determines the direction and iteratively learns the global structure. However, this method may fail under the large noise regime.

In this paper, we propose a framework named Residual Independence Test (RIT) for gradient-based DAG learning algorithms. Our work relies on the structure asymmetry, between pairing variables under proper conditions. We perform regressions on both directions between pairing variables, then test the dependence of regression residual and potential direct cause. Based on structure asymmetry, the residual and direct cause are independent for the true direction but not for the opposite direction. Hilbert Schmidt Independence Criterion (HSIC) is used as the statistics test metric [22], which is suitable for parametric and non-parametric distribution. Our framework is flexible for both continuous optimization and neural network model for DAG learning. We conduct extensive experiments on both synthetic and benchmark Bayesian network datasets. The results show that our framework has significantly improved the performance of the state-of-the-art gradient-based approaches for DAG learning.

The remaining part of this paper is structured as follows. Section 2 reviews existing work of DAG learning. Section 3 gives a problem definition and preliminary knowledge. Section 4 presents our proposed framework. Section 5 shows the experiment results. In Section 6, we conclude our work.

II. Related Work

In this section, we briefly review work relevant to DAG learning. Traditional algorithms of DAG learning can be roughly divided into two categories: constraint-based and score-based methods. Based on Markov and faithful assumptions, classical constraint methods such as Peter-Clark (PC) and Fast Causal Inference (FCI) [23], firstly employ conditional independence tests to determine the skeleton of underlying causal structure, then extend it to a Completed Partially Directed Acyclic Graph (CPDAG) by three orientation rules [23]. Differently, FCI can tolerate and even discover the existence of agnostic confounding variables. However, conditional independence tests for constraint-based methods would be intractable if the distribution of data is unknown. And the faithfulness is a so strong assumption that conditional independence tests would easily introduce errors to the recovered DAG in the case of limited samples.

Score-based methods assign a score to each candidate DAG according to some predefined score functions and adopt a greedy search strategy to identify a high-scoring one. For example, given a predefined score function called Bayesian Dirichlet equivalence uniform (BDeu) [24], Greedy Equivalence Search (GES) [25] begins with a completely empty graph and searches the optimal DAG by adding, removing or reversing edges. Though score-based methods generally learn the optimal DAG in the case of infinite variables, the time complexity increases exponentially with the number of variables. Subsequently, the score-based algorithm has produced some variants such as Greedy Interventional Equivalence Search (GIES) [26] and bnlearn [27]. However, when faced with large-scale practical problems, these score-based methods usually need to add extra structural assumptions. To overcome this issue, a hybrid algorithm named MMHC [19] has been proposed, which firstly performs conditional independence tests to build an initial skeleton and then adopts a score function to identify the direction of each edge in the learnt skeleton.

As discussed above, constraint-based and score-based methods obtain only the Markov equivalence class of the true DAG. In contrast, a series of algorithms based on the Functional Causal Model (FCM) or Structural Equation Model (SEM) can learn a complete causal graph. For example, LiNGAM [28] has shown that the underlying DAG can be fully identified under the condition of non-Gaussian noise, linearity and causal sufficiency assumptions. Additive Noise Model (ANM) [29] proves that the linear-non-Gaussian FCM can be generalized to admit nonlinear dependencies as long as the noise remains additive.

Considering combinatorial explosion problem, traditional DAG learning methods usually deal with discrete data [30]. Recently, NOTEARS [16] converts the combinational optimization problem into a continuous optimization program with an acyclicity constraint. NOTEARS utilizes gradient descent to determine the DAG in the form of a weighted adjacency matrix and achieves good structure recovery results in the case of linear FCM. Further, to avoid the noise of variables be absorbed into the causal graph reconstruction model, DARING [31] adds an extra adversarial network to implement residual independence constraint. NOTEARS-MLP [17] adopts neural network and orthogonal basis expansion to fit the generative process of separative variable so that it can handle more complex nonlinear causal mechanisms. DAG-GNN [18] combines variational autoencoder [32] with DAG learning and adopts Evidence Lower Bound (ELBO) as the objective function. GAE [33] uses a model frame like DAG-GNN but abandons variation inference, ultimately increases the accuracy while shortening the time, particularly in large-scale variables. SAM [34] learns causal graph by a generative adversarial network [35], which defines a causal graph as binary structure gates of the generative network and adds a smooth acyclicity constraint to the objective function. Gran-DAG [14] and MaskedNN [15] use neural networks to approximate the underlying data generating functions and equivalently define the weighted adjacency matrix by path products of neural network. For all these methods, the causal structure is defined as a weighted adjacency matrix optimized by the variations of gradient descent.

III. Background

A. Problem Definition

Let $G \in \mathbb{R}^{d \times d}$ be a DAG with d variables $X = \{X_1, X_2, ..., X_d\}$. For $X_i \in X$, we define $X_{pa(i)}$ to denote the parent set of variable X_i so that there is an edge from $X_j \in X_{pa(i)}$ to X_i. A commonly used causal mechanism model in DAG learning is the Functional Causal Model (FCM) [36], which makes each node can be generated from a function of its parent nodes and a unique noise. If the noise variables are jointly independent and satisfy the additive condition, such model is also called ANM [29], that is,

$$X_i = f_i(X_{pa(i)}) + Z_i \tag{1}$$

where f_i is the data generative function for variable X_i and Z_i is an external noise.

In this paper, we suppose that the data generative process follows ANM, our purpose is to identify the true graph from observational data $X = \{X^{(i)}\}_{i=1}^{n}$ under proper conditions.

B. Structural Asymmetry

In practice, there is no way to eliminate local uncertainties that joint distribution allows the corresponding FCM to indicate either $X_i \to X_j$ or $X_j \to X_i$ in the causal graph [37]. This problem becomes more troublesome when discovering DAG through gradient-based approaches. From a global perspective, gradient-based methods including continuous optimization and neural network, estimate the weighted adjacency matrix in the form of parameters, which are easy to overlook the local details. Beyond that, these methods mainly focus on reducing the least square loss between real samples and reconstructed samples or the maximum likelihood estimation of reconstructed samples. This idea is beautiful in theory, but its practical effect is subject to many factors like the structures and parameters of the neural network.

Actually, given some proper assumptions about the functional and parametric form of the data generative process, one can adopt structural asymmetry to determine the correct direction of an edge [36] [37]. Reviewing the ANM, we assuming that the data generative process follows the linear causal mechanism $X = U_X$, $Y = X + U_Y$ and U_Y is independent of X. We expect the regression residual of the true direction reflects the property $U_Y \perp\!\!\!\perp X$. Actually, in the linear, non-Gaussian and acyclic model (LiNGAM), if at most one of the noise term U_Y and the cause X is Gaussian, the direction $X \to Y$ is identifiable [36].

We show this property in Fig. 2: the true structural equation is $X = U_X$ and $Y = X + U_Y$, the left column shows the scatters of X, Y and the regression residual of Y given X. The column on the right is corresponding to regressing X on Y. And the top two figures indicate that both cause and noise are uniformly distributed, while the pictures on the bottom represent the noise and cause follow Gaussian distribution. Obviously, the real causal direction follows the structural asymmetry but the reverse one rejects it. Although residuals exhibit significant asymmetry in the linear case, the conditions are too rigorous. Recently, a more universal model with mild conditions has been proposed. The post-non-linear additive noise model (PNL) [38] supposes that the data generative process taking the form of $X = U_X$ and $Y = f(X) + U_Y$ where f is a sufficiently non-linear function and does not require either the cause X or noise U_Y to be non-Gaussian. Similar to the ANM, the PNL model also reveals structural asymmetry by the (in)dependence of regression residuals and direct causes. However, methods exploiting the structural asymmetry only can test edges respectively and result in intractable search spaces expanding superexponentially with the number of nodes.

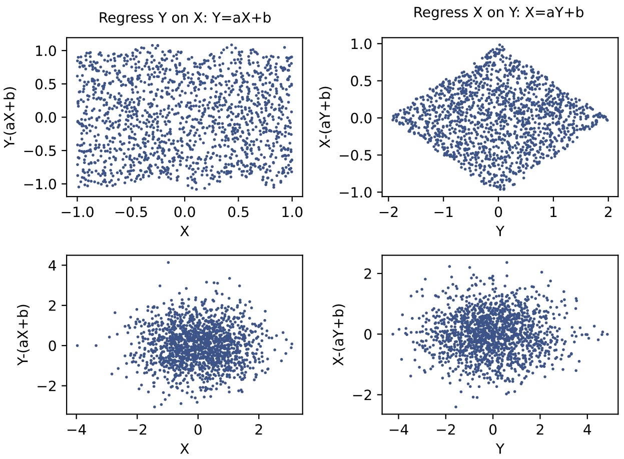


Fig. 2: Example of structural asymmetry with different noises.

IV. Our Proposed Framework

Our proposed Residual Independence Test (RIT) framework utilizes the advantages of both gradient-based DAG learning methods and structural asymmetry to tackle the drawbacks of gradient-based DAG learning methods. The main idea of the RIT framework is as follows.

Firstly, we exploit the powerful fitting ability of the neural network to identify a directed graph as a skeleton. Then, for each edge in the skeleton, we perform regressions for both directions and then execute the independence tests between the regression residuals and the hypothetical causes. The direction gives an independent regression residual would be considered as the real causal relationship. Generally, our framework is suitable for all gradient-based DAG learning methods. Additionally, since these methods cannot absolutely guarantee acyclic, our method has the extra advantage of removing circles that might be present in the skeleton.

The RIT framework employs the Hilbert Schmidt Independence Criterion (HSIC) [22] (a robust non-parametric independent test statistic) to examine structural asymmetry. HSIC adopts an injective map to transform the features into Reproducing Kernel Hilbert Spaces (RKHSs).

Let P_{xy} be a joint probability distribution on $\chi \times \gamma$ and P_x, P_y is the corresponding marginal distribution of variables x and y. Given an injective map ϕ with a positive kernel function k, we can convert the variable $x \in \chi$ into a RKHS F, that is, $\phi : x \rightarrow \phi(x) \in F$. Similarly, we define the second RKHS G on Y with kernel function l and an injective map: $\psi : y \rightarrow \psi(y) \in G$. According to [39], a cross-covariance operator $C_{xy} : G \rightarrow F$ is defined as:

$$C_{xy} := E_{xy}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] \tag{2}$$

where $\mu(x) := E_x\phi(x)$, $\mu(y) := E_y\psi(y)$ and $\otimes$ is the tensor product. C_{xy} is a generalization of the cross-covariance matrix between random vectors and the largest singular value of this operator, $||C_{xy}||$, is equal to zero if and only if $x \perp\!\!\!\perp y$ [22]. Although the largest singular gives a criterion measuring dependence between variables, it requires restrictive function classes. According to [22], a more general description of HSIC is to define it as the squared Hilbert-Schmidt norm of the cross-covariance operator:

$$\begin{aligned} HSIC(P_{xy},F,G) &= ||C_{xy}||^2 \\ &= E_{xx'yy'}[k(x,x')l(y,y')] \\ &+ E_{xx'}[k(x,x')]E_{yy'}[l(y,y')] \\ &- 2E_{xy}[E'_x[k(x,x')]E'_y[l(y,y'))]] \end{aligned} \tag{3}$$

where x' denotes an independent copy of $x \in \chi$, and the positive kernel function we used is $k(x,x') = exp(-\frac{||x-x'||^2}{\sigma^2})$. Thus, given m samples $(x,y) = \{(x_1,x_2),...,(x_m,y_m)\}$ drawn independently from P_{xy}, then $P_{xy} = P_xP_y$ if and only $HSIC(P_{xy},F,G) = 0$, where F and G are two RKHSs related to x and y.

Now, we present the implementation process of our proposed framework. As shown in algorithm 1, our framework can be divided into Phase 1 and Phase 2.

Algorithm 1: The RIT framework

```
Input: i.i.d samples $X = \{X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}\}_{i=1}^n$, threshold $\varepsilon$,
    maximum number of iterations $T$
Output: causal structure $G$
 1: // Phase 1: Learn a skeleton $G$ by gradient-based
    methods
 2: Initialize $G$ and parameters $\theta$ of causal models
 3: repeat
 4:     generate n samples $X = \{X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}\}_{i=1}^n$
 5:     compute the objective function $L_c(G,\theta,\rho)$
 6:     update $G$, $\theta$ to optimize $L_c(G,\theta,\rho)$
 7: until arrive maximal iteration number $T$ or trigger
    termination conditions;
 8: prune the edges less than $\varepsilon$ in $G$
 9: // Phase 2: Correct the reverse edges in the skeleton $G$
10: for each $(x,y)$ in $G$ do
11:     regress $y$ on $x$ and compute $R_{xy} = y - \hat{f}_y(x)$
12:     calculate independence criteria $H_{xy} = HSIC(R_{xy},x)$
13:     regress $x$ on $y$ and compute $R_{yx} = x - \hat{f}_x(y)$
14:     calculate independence criteria $H_{yx} = HSIC(R_{yx},y)$
15:     compute the score $H = H_{yx} - H_{xy}$
16:     if $H > 0$ then
17:         no operation for the true direction
18:     else
19:         $(x,y) \leftarrow (y,x)$
20:         assert no circle in $G$
21:     end if
22: end for
23: return $G$
```

Phase 1 (lines 2 to 8): we employ existing gradient-based methods to learn a skeleton of the underlying DAG. Firstly, we adopt SEM and its variants to generate reconstruction samples given i.i.d samples $X = \{X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}\}_{i=1}^n$. Then the DAG learning problem is formulated as an optimization problem with an objective function including the evaluation of data reconstructions, sparsity, and smooth acyclicity constraint. Specifically, the model evaluation criterion would be a least square loss between real samples and reconstructed samples. Sparse constraint is generally represented by $||G||_1$. To ensure the acyclic characteristic, a smooth acyclic constraint is proposed for continuous optimization [16]:

$$h(G) = tr(e^{G \odot G}) - d = 0 \tag{4}$$

where $\odot$ denotes the Hadamard product and $e^M = \sum_{k=0}^{+\infty} \frac{M^k}{k!}$ is the power series expansion [16]. Then $(M^k)_{ij}$ is the sum of weight products along all k-step paths from node j to node i. Thus, $tr(e^{G \odot G}) - d = 0$ equivalently represents that there is no cycle in G. Then, the DAG learning problem is converted to optimize the following equality-constraint (ECP) program [33]:

$$\min_{A \in R^{(d \times d)}} L(G,\theta) = \frac{1}{2n} \sum_{j=1}^{n} ||X^{(j)} - G^T X^{(j)}||_2^2 + \lambda ||G||_1 \tag{5}$$

$$\text{subject to } h(G) = tr(e^{G \odot G}) - d = 0$$

This ECP problem can be solved by augmented lagrangian method [33]:

$$L_c(G,\theta,\rho) = L(G,\theta) + \rho h(G) + \frac{c}{2}|h(G)|^2 \tag{6}$$

where θ is the parameter of model, ρ is the Lagrange multipier and c is the penalty parameter. Then we have the following parameters updating rules [18]:

$$(G^k, \theta^k) = \underset{G,\theta}{\arg\min} L_{c^k}(G,\theta,\rho^k), \tag{7}$$

$$\rho^{k+1} = \rho^k + c^k h(G^K), \tag{8}$$

$$c^{k+1} = \begin{cases} \eta c^k, \ if |h(G^k)| > \gamma |h(G^{k-1})|, \\ c^k, \ otherwise. \end{cases} \tag{9}$$

where $\eta > 1$ and $\gamma < 1$ are tuning parameters [18]. After updating the parameters iteratively, a skeleton G is learnt. Finally, we prune the entries in the skeleton G where their absolute value is less than threshold $\varepsilon = 0.3$.

Phase 2 (lines 10 to 22): we correct the reversed edges by calculating the dependence of regression residuals and directed causes. For every edge $x \rightarrow y$ in the graph obtained from phase 1, we do a regression on direction $x \rightarrow y$, $y \rightarrow x$ and

TABLE I: Introduction of benchmark BNs.

Network	Num. Vars	Num. Edges	Max In/Out-Degree	Min/Max |PCset|
alarm	37	46	4/5	1/6
barley	48	84	4/5	1/8
carpo	74	81	5/12	0/12
hailfinder	56	66	4/16	1/17
mildew	35	46	3/3	1/5

compute the corresponding residual R_{xy} and R_{yx}. Then we calculate the dependence between the regression residuals and the hypothetical causes, i.e., H_{xy} and H_{yx}, respectively. Finally, we define a score $H = H_{yx} - H_{xy}$ to indicate the direction of the current edge. For the true direction $x \to y$, $R_{xy} \perp\!\!\!\perp x$, which means the value of $HSIC(R_{xy}, x)$ is approaching to 0. While $H_{yx} = HSIC(R_{yx}, y)$ is a positive real number due to the fact that $R_{yx} \not\!\perp\!\!\!\perp y$ means the value of HSIC is greater than 0. We do no operation for this case. But when $H < 0$, we reverse the direction of the current edge and assert this process would not introduce a circle to G.

V. Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our proposed framework. The structure of this section is organized as follows. Section 5.1 introduces the experiment settings. Section 5.2 shows the results of our proposed methods and three baseline algorithms on linear datasets. Section 5.3 gives the results on the non-linear datasets. Section 5.4 compares our proposed method with other DAG learning algorithms on a real dataset.

A. Experiment setting

1) Datasets: Five benchmark Bayesian networks (BNs) are used to evaluate the performance of our proposed framework. The details of the benchmark BNs are summarized in Table I.

The synthetic datasets are generated in the manner of following causal mechanism:

- *linear:* X are generated from the linear SEM $X = A^T X + Z$, where the non-zero terms in the coefficient matrix are sampled from $N(0,1)$ and the noise Z comes from $U(-2,2)$.
- *non-linear:* X are generated from sufficiently non-linear function $X = A^T cos(X+1) + Z$, the parameter settings of the nonlinear case are consistent with the linear case except the external noise is a mixture of $N(0,1)$ and $U(-2,2)$.

For each benchmark BN, 5 samples with sizes of 500, 1000 and 5000 were generated according to the topological order of the graph, respectively.

2) Implementation Details: Based on NOTEARS [16], NOTEARS-MLP [17], and DAG-GNN [18], we instantiate the RIT framework and generate three corresponding methods, NOTEARS+RIT, NOTEARS-MLP+RIT, and DAG-GNN+RIT. Then we compare these three methods with NOTEARS, NOTEARS-MLP, and DAG-GNN on synthetic datasets. For all these original algorithms, we directly used the source codes provided by the authors. Meanwhile, we compare our proposed NOTEARS+RIT, NOTEARS-MLP+RIT, and DAG-GNN+RIT with PC [23], LiNGAM [28], bnlearn [27], GIES [26], SAM [34] and ANM [29] on a real dataset. The implementation of these comparative algorithms is available at https://fentechsolutions.github.io/CausalDiscoveryToolbox. We adopt 0.3 as the threshold ε to prune the DAGs obtained from gradient-based methods.

3) Evaluation Metrics: Assuming that TP is the number of true positive items representing that an edge in the ground truth DAG is correctly found through the algorithm. And FP is the number of false positive items indicating that there exiting an edge in the predicted DAG but not in the true DAG. TN is the number of edges that do not exist in true DAG and the predicted DAG. FN is the number of edges existing in the true DAG but missing in the predicted DAG. We evaluate the efficiency of our frameworks by the following metrics.

- *False Discovery Rate (FDR).* FDR is the expected proportion of type I errors. Equivalently, it is the expected ratio of the number of false positive classifications to the total number of positive classifications: $FDR = \frac{FP}{TP+FP}$.
- *True Positive Rate (TPR).* TPR is also called recall, which refers to the probability that an actual positive will test positive. TPR is calculated as: $TPR = \frac{TP}{TP+FN}$.
- *False Positive Rate (FPR).* In statistics, FPR is the probability of falsely rejecting the null hypothesis for a particular test. FPR is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events: $FPR = \frac{FP}{FP+TN}$.
- *Structural Hamming Distance (SHD).* SHD is an effective metric for measuring the difference between the found graph and the ground truth graph. It is the numbers of the extra edges, reverse edges and missing edges in the found graph.

Note that in the following experiments, the lower values of SHD, FDR, and FPR mean better performance of an algorithm, while the higher values of TPR means better performance of an algorithm.

B. Results of DAG learning on linear synthetic data

In this section, we compare the effectiveness of our proposed framework with the original algorithms on five synthetic datasets. From Fig.3, we can see that:

- For all datasets, NOTEARS and DAG-GNN have similar performance. But we notice that NOTEARS and DAG-GNN perform better than NOTEARS-MLP on almost all datasets, which indicates that NOTEARS-MLP does not model true causality between variables well.
- No matter based on NOTEARS, NOTEARS-MLP or DAG-GNN, the implementation of our proposed framework consistently improves their performance on four metrics. And the significant promotion of TPR and reduction of SHD, FDR implies our framework can accurately

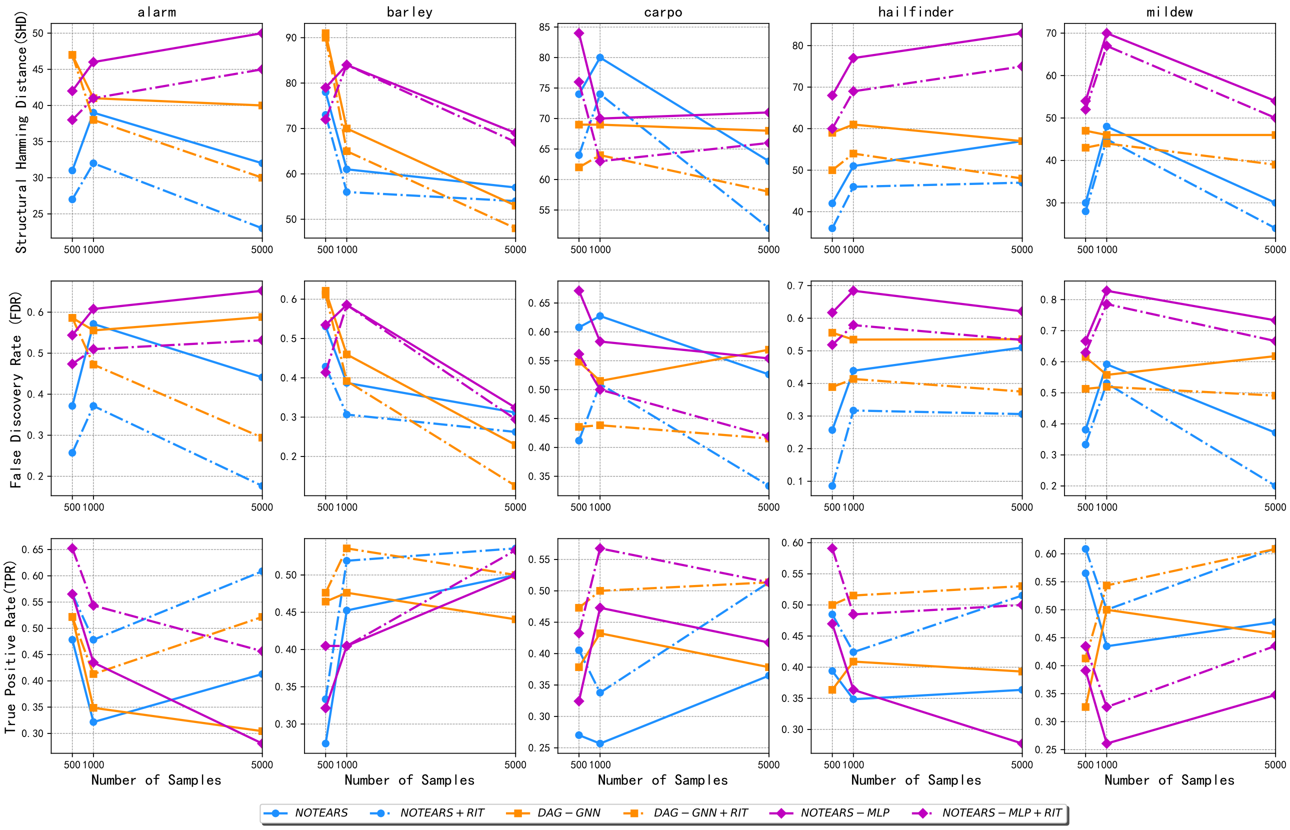


Fig. 3: Results of DAG learning algorithms on linear datasets.

find the reverse directions and convert them to the true directions.

- The little improvement on barley can be explained by the low FDR, which means there exist few reverse edges in the primal causal graph.

C. Results of DAG learning on nonlinear synthetic data

To demonstrate the effectiveness of our proposed framework on nonlinear datasets, we conduct the same experiments. From the empirical results reported in Fig 4, we can see that:

- We notice that NOTEARS and DAG-GNN perform poorly on nonlinear datasets because their models are based on linear SEM. However, NOTEARS-MLP is specially designed for nonlinear causal mechanisms so that it is relatively accurate in the nonlinear case.
- Since we have used a mixture of Gaussian noise and uniform noise in the nonlinear case, which brings interference to the data generation processes. At this point, the performance of NOTEARS, DAG-GNN and NOTEARS-MLP becomes poor in practice. More specifically, the TPR is greatly reduced and the FDR is significantly promoted.
- In the nonlinear case, our framework can also promote the accuracy of the baseline algorithm NOTEARS, DAG-GNN, NOTEARS-MLP. This shows that our framework improves the applicability of the gradient-based structure learning algorithms in different situations.

D. Results of DAG learning on real data

To compare the performance of our proposed framework on a real dataset, we consider a real bioinformatics dataset Sachs [40]. Sachs is a protein signaling network expressing the level of different proteins and phospholipids in human cells. It is commonly viewed as a benchmark graphical model with 11 nodes and 17 edges. In our experiments, we adopt the observational data with 853 samples.

In Table II, we compare our proposed three algorithms with PC, GIES, ANM, SAM, bnlearn, LiNGAM, NOTEARS, DAG-GNN, and NOTEARS-MLP. From the results, we can see that the performance of NOTEARS and DAG-GNN is poor. It can be explained by their linear SEM, which is not suit-

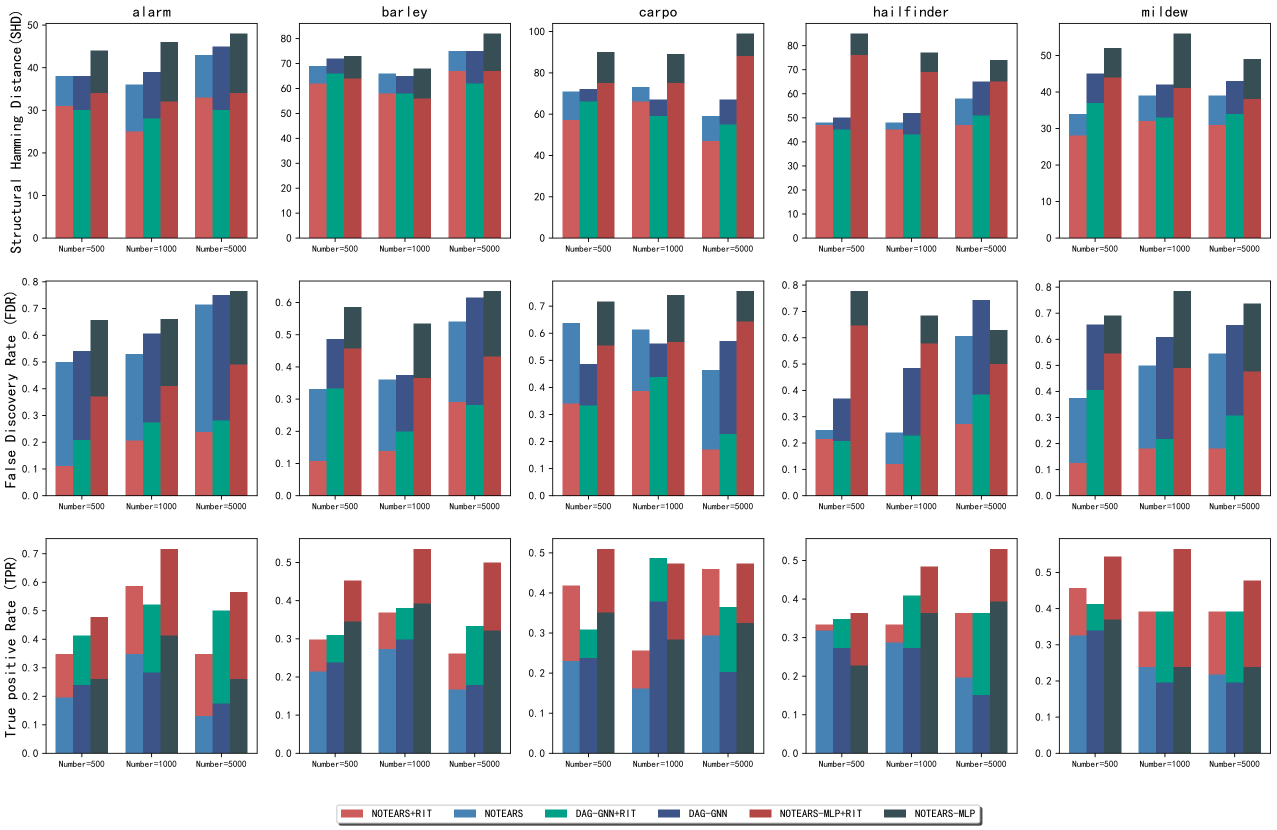


Fig. 4: Results of DAG learning algorithms on nonlinear datasets.

TABLE II: Results on the real Sachs dataset (↓ means that the lower, the better while ↑ represents the higher, the better.)

Algorithms	SHD(↓)	Reverse edges(↓)	FDR(↓)	TPR(↑)	FPR(↓)
PC	21	1	0.6667	0.3529	0.4473
GIES	25	2	0.7619	0.2941	0.4210
ANM	17	3	0.5882	0.4117	0.2632
SAM	33	3	0.7895	0.4706	0.7894
bnlearn	23	1	0.7391	0.3529	0.3158
LiNGAM	16	3	0.6250	0.1765	0.1316
NOTEARS	19	7	0.8000	0.1764	0.3158
NOTEARS+RIT	15	3	0.5333	0.4118	0.2105
DAG-GNN	22	7	0.8333	0.1764	0.3947
DAG-GNN+RIT	19	4	0.6667	0.3529	0.3158
NOTEARS-MLP	20	4	0.8181	0.1176	0.2380
NOTEARS-MLP+RIT	17	1	0.5454	0.2941	0.1579

able for complex real data. However, our proposed framework significantly improves the performance of NOTEARS, DAG-GNN and NOTEARS-MLP and this validates that our methods can correct the reversed edges. GIES, ANM and LiNGAM also show competitive results on the metric SHD and reversed edges. PC, bnlearn and NOTEARS-MLP+RIT achieve the best performance in terms of the number of reversed edges.

VI. Conclusion

In this paper, we propose a framework utilizing the advantage of both gradient-based DAG learning methods and structural asymmetry to tackle the shortcomings of existing gradient-based DAG learning methods. This framework can uncover the real direction of an edge by testing the (in)dependence of the regression residual and cause. The experimental results have demonstrated that our framework significantly improves the performance of DAG learning on both synthesis and real datasets.

Acknowledgments

This work is supported by National Natural Science Foundation of China (under Grant 61876206) and the Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province (under grant CICIP2020003).

References

[1] S. Yang, H. Wang, K. Yu, F. Cao, and X. Wu, “Towards efficient local causal structure learning,” *IEEE Transactions on Big Data*, 10.1109/TBDATA.2021.3062937, 2021.

[2] Z. Ling, K. Yu, H. Wang, L. Li, and X. Wu, "Using feature selection for local causal structure learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 4, pp. 530–540, 2021.

[3] M. J. Vowels, N. C. Camgoz, and R. Bowden, "Targeted vae: Structured inference and targeted learning for causal parameter estimation," *arXiv preprint arXiv:2009.13472*, 2020.

[4] K. Yu, M. Cai, X. Wu, L. Liu, and J. Li, "Multilabel feature selection: A local causal structure learning approach," *IEEE Transactions on Neural Networks and Learning Systems*, 10.1109/TNNLS.2021.3111288,2021.

[5] K. Yu, L. Liu, J. Li, W. Ding, and T. D. Le, "Multi-source causal feature selection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2240–2256, 2019.

[6] B. Schölkopf, "Causality for machine learning," *arXiv preprint arXiv:1911.10500*, 2019.

[7] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, and R. van Lier, *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018.

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

[9] M. Tsagris, "Bayesian network learning with the pc algorithm: an improved and correct variation," *Applied Artificial Intelligence*, vol. 33, no. 2, pp. 101–123, 2019.

[10] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–36, 2020.

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

[12] D. M. Chickering, "Learning equivalence classes of bayesian-network structures," *The Journal of Machine Learning Research*, vol. 2, pp. 445–498, 2002.

[13] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1551–1560.

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

[15] I. Ng, Z. Fang, S. Zhu, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," *arXiv preprint arXiv:1910.08527*, 2019.

[16] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *arXiv preprint arXiv:1803.01422*, 2018.

[17] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, "Learning sparse nonparametric dags," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3414–3425.

[18] Y. Yu, J. Chen, T. Gao, and M. Yu, "Dag-gnn: Dag structure learning with graph neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7154–7163.

[19] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[20] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf, "Cause-effect inference by comparing regression errors," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 900–909.

[21] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, "Information-geometric approach to inferring causal directions," *Artificial Intelligence*, vol. 182, pp. 1–31, 2012.

[22] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola *et al.*, "A kernel statistical test of independence." in *Nips*, vol. 20. Citeseer, 2007, pp. 585–592.

[23] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[24] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[25] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[26] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2409–2464, 2012.

[27] D. Margaritis, "Learning bayesian network model structure from data," Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, Tech. Rep., 2003.

[28] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[29] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf *et al.*, "Nonlinear causal discovery with additive noise models." in *NIPS*, vol. 21. Citeseer, 2008, pp. 689–696.

[30] A. Mohammadi and E. C. Wit, "Bayesian structure learning in sparse gaussian graphical models," *Bayesian Analysis*, vol. 10, no. 1, pp. 109–138, 2015.

[31] Y. He, P. Cui, Z. Shen, R. Xu, F. Liu, and Y. Jiang, "Daring: Differentiable causal discovery with residual independence," 2021.

[32] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

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

[34] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag, "Structural agnostic modeling: Adversarial learning of causal graphs," *arXiv preprint arXiv:1803.04929*, 2018.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[36] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like dags? a survey on structure learning and causal discovery," *arXiv preprint arXiv:2103.02582*, 2021.

[37] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[38] K. Zhang and A. Hyvarinen, "On the identifiability of the post-nonlinear causal model," *arXiv preprint arXiv:1205.2599*, 2012.

[39] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *Journal of Machine Learning Research*, vol. 5, no. Jan, pp. 73–99, 2004.

[40] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.