Effective Causal Discovery under Identifiable Heteroscedastic Noise Model

Naiyu Yin[1], **Tian Gao**[2], **Yue Yu**[3], **Qiang Ji**[1]

[1]Rensselaer Polytechnic Institute, Troy, NY.
[2]IBM Research, Yorktown Heights, NY.
[3]Lehigh University, Bethlehem, PA.

Abstract

Capturing the underlying structural causal relations represented by Directed Acyclic Graphs (DAGs) has been a fundamental task in various AI disciplines. Causal DAG learning via the continuous optimization framework has recently achieved promising performance in terms of both accuracy and efficiency. However, most methods make strong assumptions of homoscedastic noise, i.e., exogenous noises have equal variances across variables, observations, or even both. The noises in real data usually violate both assumptions due to the biases introduced by different data collection processes. To address the issue of heteroscedastic noise, we introduce relaxed and implementable sufficient conditions, proving the identifiability of a general class of SEM subject to these conditions. Based on the identifiable general SEM, we propose a novel formulation for DAG learning that accounts for the variation in noise variance across variables and observations. We then propose an effective two-phase iterative DAG learning algorithm to address the increasing optimization difficulties and to learn a causal DAG from data with heteroscedastic variable noise under varying variance. We show significant empirical gains of the proposed approaches over state-of-the-art methods on both synthetic data and real data.

1 Introduction

Learning the statistical and causal dependencies of a distribution in the form of a directed acyclic graph (DAG) is of great interest in areas such as causal inference and Bayesian network structure learning. The underlying statistical or causal relations indicated by the DAG have been applied to various machine learning applications (Ott, Imoto, and Miyano 2004; Spirtes, Meek, and Richardson 1995). Causal DAG plays an increasingly important role in many machine learning tasks, including out-of-distribution generalization (Janzing and Schölkopf 2018; Shen et al. 2018; Ahuja et al. 2021), domain adaptation (Javidian, Pandey, and Jamshidi; Stojanov et al. 2021), and transfer learning (Schölkopf 2019).

The gold standard approach to performing causal discovery is to conduct controlled experiments, which can be expensive, time-consuming, and sometimes even infeasible. Therefore, algorithms have been proposed to learn a DAG from purely observational data. These algorithms can be divided into two categories: constraint-based methods and score-based methods. The constraint-based methods estimate DAGs by performing independence tests between variables. Popular algorithms include PC (Spirtes et al. 2000) and FCI (Spirtes, Meek, and Richardson 1995; Zhang 2008). The score-based methods search through the DAG space for a DAG with the optimal score. The differences among score-based methods usually come from search procedures, such as hill-climbing and Greedy Equivalent Search (GES) (Chickering 2002). The structural causal model-based methods encode the statistical and causal dependencies via structural equation models (SEM). Zheng et al. (2018) introduces a continuous DAG constraint and NOTEARS algorithm, which reformulates the original combinatorial DAG learning problem as a constrained continuous optimization. Such conversion enables the employment of continuous optimization techniques in follow-up works (Kalainathan et al. 2018; Yu et al. 2019; Ng et al. 2019).

Under either a linear or non-linear structural equation model (SEM) assumption, most of the current methods (Zheng et al. 2018, 2020; Yu and Gao 2020; Peters et al. 2014) usually adopt an assumption in SEM that the noises are additive to causal functions and are assumed to have equal variance for each variable. However, such an assumption may not hold in real-world data. For example, real-world data may be gathered from diverse sources, spanning different times and locations, employing a variety of collection techniques. As a result, the exogenous factors that impact each variable may differ, and noise variances become non-constant for observations. Incorrect assumptions regarding variable noise homoscedasticity, when they are heteroscedastic, may lead to inaccurate and biased estimates. Several works (Ng, Ghassami, and Zhang 2020; Lachapelle et al. 2019; Park 2020) seek to allow the noises of each variable to have different variances but fall short of fully addressing noise heteroscedasticity.

A few recent works explicitly extend the SEM with additive noise assumption to more general cases and estimate the noise observation heteroscedasticity. Rajendran et al. (2021) employs SEM with multiplicative noise, while Blöbaum et al. (2018) assumes the existence of a joint distribution between noise and parent variables. Lachapelle et al. (2019), Xu et al. (2022); Immer et al. (2022), Khemakhem et al.

(2021), Duong and Nguyen (2023) modulate the noise variances as a deterministic function of the parent variables. However, these works adopt bivariate SEMs and infer pairwise causal relations. To learn a causal DAG with more than two variables, they need to estimate the causal order or the skeleton first using existing methods.

To accurately estimate the DAG from data with heteroscedastic variable noises and varying residual variance across observations, we propose employing a more general form of SEM and, thus, designing a novel DAG structure learning formulation. The main advantage of using a general SEM lies in relaxing the assumptions on noise variances, allowing not only unequal variances across variables but also varying variances across observations for the same variables. Such relaxation reduces model misspecification and enables the algorithm to more accurately capture noise variances and learn DAGs from challenging yet realistic data. However, this relaxation also significantly increases the difficulties in optimization modeling (Lachapelle et al. 2019).

Main Contributions: To tackle those issues, we make three major contributions: 1) We introduce relaxed, implementable sufficient conditions for the identifiability of a general class of multivariate SEM. Guided by the identifiability conditions, we propose a novel DAG learning formulation that considers the variability of noise variances both among variables and across observations. To achieve this, our formulation models the parameters of the noise distribution with neural networks (Eq. (9)). 2) We present an effective and practical two-phase DAG learning algorithm, which iteratively minimizes the objective to ensure accurate estimation of noise variances and DAG. 3) Empirical results demonstrate that our method achieves comparable accuracy on synthetic homoscedastic noise data compared to state-of-the-art methods. Moreover, it significantly outperforms these methods on synthetic heteroscedastic data and real data.

2 Related works

In an SEM with Gaussian additive noise, functional causal model-based methods, such as Chen, Drton, and Wang (2019), assume the variables have homoscedastic noises[1] with equal noise variances across observations. In other words, the variable noises have equal variance across both variables and observations. The strong homoscedastic assumption is also implicitly posed for methods (Zheng et al. 2018, 2020; Yu et al. 2019; Gao, Ding, and Aragam 2020) that adopt reconstruction loss under the same SEM setting[2]. Ng, Ghassami, and Zhang (2020) relaxes the homoscedastic variable noise assumption, allowing the noises of different variables to have non-equal variances. Similarly, Lachapelle et al. (2019) and Park (2020) perform the same relaxation.

Moreover, the above methods assume equal noise variances for each variable across observations, whereby the variable noise variance may vary from observation to observation due to the variation of the data collection conditions. Noise observation heteroscedasticity modeling has received increasing attention over the past few years. The general approach is to relax the independence between the parent variables and the additive noise. Blöbaum et al. (2018) allows a dependency between parent variables and the noise by assuming a joint distribution of two terms exists. Xu et al. (2022) models the noise variance as a piece-wise function of the parent variables with limited choices of variance values. Khemakhem et al. (2021); Immer et al. (2022); Duong and Nguyen (2023) employ a general form of SEM and modulate the noise variance as a deterministic function of the parent variables. However, Khemakhem et al. (2021) and Immer et al. (2022) are mainly designed to identify pair-wise cause-effect relations for bivariate SEMs. Duong and Nguyen (2023) proposes to estimate the causal order and then orient the pair-wise causal directions for multivariate SEMs. An extension of GraN-DAG, denoted as GraN-DAG++, also estimates the noise variances as a function of parent variables and learns a DAG for the multivariate case. However, due to the heteroscedasticity complexity and optimization limitation, GraN-DAG++ learns at best comparable accurate DAG. Rajendran et al. (2021) employs the multiplicative SEMs to model the heteroscedastic noise data but learn the causal structure via a discrete optimization framework. We summarize the above methods in Table 1.

In the following section, we first introduce the general form of SEM. Then we introduce sufficient conditions that provide theoretical justification for its identifiability on multivariate variables. We then propose a general DAG learning formulation, which cannot only accurately model the variation of noise variance across both variables and observations but also capture a more accurate DAG structure in complex and noisy real-world datasets or applications.

3 Background and formulation

3.1 Preliminaries

Structural Equation Model (SEM) with additive noise: Let X be a set of N random variables, $X = [X_1, X_2, \cdots, X_N]$. The causal relations between a variable $X_n \in X$ and its parents can be modeled via Eq. (1):

$$X_n = f_n(X_{\pi_n}) + E_n, n = 1, 2, \cdots, N \tag{1}$$

where $f_n(\cdot)$ is the structural causal function. X_{π_n} are the parent variables of X_n. E_n is the exogenous noise variable corresponding to variable X_n. Together they account for the effects from all the unobserved latent variables and are assumed to be mutually independent (Peters, Janzing, and Scholkopf 2011).

DAG structure learning under SEM: To learn a DAG $\mathcal{G}$ from a given joint distribution $P(X)$, X is usually modeled via SEMs defined by a set of continuous parameters $A = (A_1, A_2, \cdots, A_N)$ that encode all the causal relations, i.e.,

$$X_n = f_n(X; A_n) + E_n, n = 1, 2, \cdots, N \tag{2}$$

where A_n are the parameters in each SEM. Compared to Eq. (1), it is easy to see that A_n selects parent variables X_{π_n} for each X_n. The goal is to estimate A, based on which we can infer the DAG $\mathcal{G}$. Let $\mathbf{X} \in \mathbb{R}^{M \times N}$ denote the input matrix of M observations of the random variable set

[1] If a set of variable noises is homoscedastic, then they have equal variances.

[2] Please refer to supplementary section 4 for details.

Table 1: Summary of SEMs and algorithms for SoTA methods.

SoTA Methods	SEM				Algorithm
	# var.	Causal function	Noise	Identifiable	Optimization
NOTEARS (Zheng et al. 2018)	Multivariate	Linear	Homoscedastic	✓	Continuous
NOTEARS-MLP (Zheng et al. 2020)	Multivariate	Nonlinear	Homoscedastic	✓	Continuous
GOLEM-EV/NV (Ng, Ghassami, and Zhang 2020)	Multivariate	Linear	Homoscedastic	✗	Continuous
GraN-DAG (Lachapelle et al. 2019)	Multivariate	Nonlinear	Homoscedastic	✓	Continuous
GraN-DAG++ (Lachapelle et al. 2019)	Multivariate	Nonlinear	Heteroscedastic	✗	Continuous
US(Park 2020)	Multivariate	Linear	Heteroscedastic	✓	Combinatorial
HEC (Xu et al. 2022)	Bivariate	Nonlinear	Heteroscedastic	✓	Combinatorial
CAFEL (Khemakhem et al. 2021)	Bivariate	Nonlinear	Heteroscedastic	✓	Combinatorial
LOCI (Immer et al. 2022)	Bivariate	Nonlinear	Heteroscedastic	✓	Combinatorial
GFBS (Gao, Ding, and Aragam 2020)	Multivariate	Linear/Nonlinear	Heteroscedastic	- (multiple SEMs)	Combinatorial
HOST(Duong and Nguyen 2023)	Multivariate	Nonlinear	Heteroscedastic	✓	Combinatorial
ICDH(Ours)	Multivariate	Nonlinear	Heteroscedastic	✓	Continuous

X. Given $\mathbf{X}$, A is estimated by minimizing the loss function $F(\mathbf{X}, A)$, subject to the continuous acyclicity constraint $h(A) = tr(e^{A \circ A}) - N = 0$ (Zheng et al. 2018, 2020)[3]

$$A^* = \arg\min_A F(\mathbf{X}, A) \qquad \text{subject to } h(A) = 0 \tag{3}$$

where $F(\mathbf{X}, A)$ evaluates the negative log-likelihood of A as the underlying relations encoded in $\mathbf{X}$. The parameterization with A, along with the introduction of continuous acyclicity constraint, transforms the DAG learning under SEM into a continuous optimization problem and enables the usage of powerful optimization techniques.

General SEM and identifiability issue. To ensure the employed SEMs are identifiable, i.e., a unique graph $\mathcal{G}$ can be identified from the joint distribution $P(X)$ generated from SEMs, the exogenous variable is usually assumed to be additive (Eq. (1)). The general SEMs in Eq. (4) are proven to be unidentifiable without any constraint (Zhang, Zhang, and Schölkopf 2015).

$$X_n = f_n(X_{\pi_n}, E_n), n = 1, 2, \cdots, N \tag{4}$$

However, some recent works try to investigate the identifiability of the general SEM with weaker assumptions and develop DAG learning methods based on the identifiable SEM.

3.2 Problem statement

We first introduce one of the general SEM formulations in **Definition 3.1** that modulate the noise variance with cause variables. The SEMs we consider in **Definition 3.1** are about SEMs with heteroscedastic additive noise. It generalizes the causal function $f_n(\cdot)$ in Eq. (2) from merely an additive transformation of causes and exogenous noise to both affine and additive transformation. Such generalization increases the model's ability to approximate data with more complex types of noise. The general SEM can address data heteroscedasicity, whereby the noise variances vary across variables and observations, depending on the causes.

Definition 3.1. *(Heteroscedastic noise model) The SEMs are heteroscedastic noise models (HNMs) if Eq.* (5) *holds for each $X_n \in X$,*

$$X_n = f_n(X_{\pi_n}) + \sigma_n(X_{\pi_n})E_n, n = 1, 2, \cdots, N \tag{5}$$

where $E_1, E_2, \cdots, E_n$ are statistically independent and all follow Gaussian distributions. $\sigma_n(X_{\pi_n}) > 0$.

The investigation of DAG learning methods under HNM has been increasingly studied due to its flexibility in modeling more complex and general data generation processes in realistic data. Let $\mathbb{E}[E_n|X_{\pi_n}] = 0$ and $Var[E_n|X_{\pi_n}] = 1$, then the conditional distribution under HNM $p(X_n|X_{\pi_n}) \sim \mathcal{N}\big(f_n(X_{\pi_n}), \sigma_n^2(X_{\pi_n})\big)$.

Advantages of HNM: We choose the SEM that modulates noise variances with cause variables for three reasons. First, it relaxes the strong independence assumption between exogenous variables and observed variables. Secondly, it satisfies the assumed data generation process, whereby observations for each variable are generated using their cause variables. Moreover, it is easy to implement via deep neural networks, which are known for their ability to modeling complex data distributions.

Limitations of prior works under HNM: Xu et al. (2022) models the variance σ_n as a deterministic piece-wise function of the parent variables, which limits the approximation of the variances to a few choices. Khemakhem et al. (2021) limits their choices of f to be nonlinear and invertible functions to ensure identifiability. However, this identifiable condition cannot readily be extended to multivariate cases. For the bivariate case, the invertibility of f is easily satisfied since its inputs and outputs are values of a single variable. For the multivariate cases, the input into the f_n is the parent variables X_{π_n} of variable X_n. The dimensions match only when the number of parent variables is 1. There is no guarantee that there exists an invertible function f_n for X_n. Duong and Nguyen (2023) proposes to learn the causal DAG by first searching for the causal order and orienting edges subject to the obtained order. However, its performance is susceptible to the accuracy of independence tests, which can be challenging to perform with difficult data. Early errors in order estimation can propagate to later stages of causal direction orientation, causing the algorithm to learn inaccurate causal graphs. Moreover, due to the time complexity of subset independence tests, the algorithm cannot scale up to large models.

Therefore, **our goal is to formulate the DAG learning problem under the identifiable multivariate HNM into a continuous optimization framework and solve the optimization with powerful tools such as neural networks.**

[3]The continuous DAG constraints for linear SEM and nonlinear SEM are introduced respectively in (Zheng et al. 2018) and (Zheng et al. 2020). We use $h(Z)$ to refer that the acyclicity constraint is posed on parameters Z, regardless of SEM types.

To do so, we first introduce relaxed implementable sufficient conditions that provide identifiability for multivariate HNM in section 3.3. Guided by those conditions, we propose our continuous DAG learning formulation in section 3.4.

3.3 Proposed identifiable HNM

In this section, we introduce the sufficient conditions for the HNM to uniquely identify a DAG from the given data distribution in **Theorem 3.2**. We can theoretically prove that the HNM is identifiable if those sufficient conditions hold.

Theorem 3.2. *(Identifiability) The formulation in Eq.* (5) *is identifiable if the following conditions are satisfied: 1)* $f_1, f_2, \cdots, f_N$ *are nonlinear; 2)* $\sigma_1, \sigma_2, \cdots, \sigma_N$ *are piecewise functions. 3)* $E_1, E_2, \cdots, E_N$ *are independent and follow Gaussian distributions*[4].

Please refer to supplementary [5] section 3 for all proofs.

The nonlinearity for f_n is in terms of $\forall X_j \in X_{\pi_n}$. The nonlinearity in terms of each input variable is slightly stronger than the nonlinearity in terms of the input parent set. However, it is easy to satisfy if we employ deep neural networks as f_ns because the nonlinear activation function is applied to each dimension of the inputs.

Comparison with identifiable PNL: The identifiable post-nonlinear model (PNL) in Zhang and Hyvarinen (2012) assumes the SEM between a variable Y and its cause X follows $Y = f_2(f_1(X) + N)$, where N is the independent noise. They further assume f_2 to be a fixed non-invertible function. Compare the PNL to the HNM, there exist cases that can be proved identifiable and covered by one model but not the other. Hence, it is impossible to compare the flexibility of the two models. They are developed to address the identifiability of different classes of SEMs.

3.4 Proposed formulation

To perform DAG learning under identifiable multivariate HNM, we parameterize Eq. (5) with a set of continuous parameters that enforce the formulation to satisfy the identifiability conditions. We instantiate Eq. (5) with continuous parameters A and B, where A, B are the parameters for causal functions $f = (f_1, f_2, \cdots, f_N)$ and variances estimation functions $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_N)$. Hence, the Eq. (5) can be then re-written as:

$$X_n = f_n(X, A_n) + \sigma_n(X, B_n)E_n, n = 1, 2, \cdots, N \quad (6)$$

There are three identifiability conditions to satisfy according to **Theorem 3.2**. To satisfy condition (3), we assume $E_n \sim \mathcal{N}(0, 1)$ for $n = 1, 2, \cdots, N$. Then we adopt 2-layer Multilayer Perceptrons (MLPs) for $f_n(\cdot)$s and $\sigma_n(\cdot)$s. By setting the activation functions as sigmoid functions for f_ns , ReLU functions for σ_ns, conditions (1) and (2) are satisfied. We use a 2-layer MLP in our formulation for simplicity. The number of layers and hidden neurons can vary as long as conditions (1) and (2) hold.

Besides the three conditions to ensure the identifiability, an underlying assumption in Eq. (6) is that the parent variables that are input into functions f_n and σ_n should be the same, or are selected from the same set. To ensure that such an assumption is always satisfied in our formulation, we design A and B to share partial parameters. In particular, we let the MLPs for f_n and σ_n share the first layer weights. We denote the first layer weights of f_n as $W_n^{(1)}$, the second layer weights as $W_n^{(2)}$, hence we have

$$f_n(X, A_n) = f_n(X, W_n^{(1)}, W_n^{(2)}) = W_n^{(2)} s(W_n^{(1)} X^T) \quad (7)$$

where $W_n^{(1)} \in \mathbb{R}^{m_1 \times N}, W_n^{(2)} \in \mathbb{R}^{1 \times m_1}$. $A_n = (W_n^{(1)}, W_n^{(2)})$. $s(\cdot)$ is the sigmoid activation function. We let σ_n share the first layer weights as f_n and denote the second layer weights for σ_n as $W_n^{(3)}$. We use a scalar parameter $W_{n0}^{(3)}$ to ensure the strict positivity of σ_n. Hence we have

$$\begin{aligned} \sigma_n(X, B_n) =& \sigma_n(X, W_n^{(1)}, W_n^{(3)}, W_{n0}^{(3)}) \\ =& \text{ReLU}\big(W_n^{(3)} s(W_n^{(1)} X^T)\big) + e^{W_{n0}^{(3)}} \end{aligned} \quad (8)$$

where $W_n^{(3)} \in \mathbb{R}^{1 \times m_1}$. $W_{n0}^{(3)} \in \mathbb{R}$. $B_n = (W_n^{(1)}, W_n^{(3)}, W_{n0}^{(3)})$. We place the acyclicity constraint on the shared parameters $W^{(1)} = (W_1^{(1)}, W_2^{(1)}, \cdots, W_N^{(1)})$ to enforce the $W^{(1)}$ to encode causal relations. Intuitively, we assume there is one unique $\mathcal{G}$, represented by the weighted matrices $W^{(1)}$. $W^{(2)} = (W_1^{(2)}, W_2^{(2)}, \cdots, W_N^{(2)}), W^{(3)} = (W_1^{(3)}, W_{10}^{(3)}, W_2^{(3)}, W_{20}^{(3)}, \cdots, W_N^{(3)}, W_{N0}^{(3)})$ are the parameters to estimate the the mean and variance using parent sets selected by $W^{(1)}$. $W^{(2)}, W^{(3)}$ may further select subsets from the parent sets for estimation. We infer our estimation of the DAG $\mathcal{G}$ from $W^{(1)}$.

Advantages of sharing parameters: The formulation that shares $W^{(1)}$ automatically ensures that f_ns and σ_ns employ the same set of parent variables as inputs. Without parameter sharing, we need to impose additional constraint that enforces the DAG structures we inferred from f_ns and σ_ns separately to be consistent with each other. Moreover, the algorithm without parameter sharing may also suffer from increased time complexity, due to the enforcement of time-consuming acyclicity constraints on parameters from both f_ns and σ_ns.

3.5 Optimization objective and difficulties

The goal is to estimate a DAG $\mathcal{G}$, given M observations of X, i.e., input matrix $\mathbf{X} = \{\mathbf{X}(m)\}_{m=1}^{M}$. $\mathbf{X}(m) \in \mathbb{R}^{1 \times N}$ is the m^{th} observation of X. $\mathbf{X}(m) = [\mathbf{X}_1(m), \mathbf{X}_2(m), \cdots, \mathbf{X}_N(m)]$, where $\mathbf{X}_n(m)$ is the m^{th} observation of variable X_n. According to the HNM, the variance for $\mathbf{X}_n(m)$ can be modeled via $\sigma_n^2(\mathbf{X}(m), B_n)$. Since $E_n \sim \mathcal{N}(0, 1)$, given $\mathbf{X}(m)$, the conditional distribution of the m^{th} observation corresponding to variable X_n given its parent variables $\mathbf{X}_{\pi_n}(m)$, i.e. $\mathbf{X}_n(m)$, can be modeled as:

$$p(\mathbf{X}_n(m)|\mathbf{X}_{\pi_n}(m)) \sim \mathcal{N}\Big(f_n(\mathbf{X}(m), A_n), \sigma_n^2(\mathbf{X}(m), B_n)\Big) \quad (9)$$

[4] E_ns are i.i.d Gaussian is a sufficient but not necessary condition of identifiability. By assuming i.i.d. Gaussian noise, sufficient conditions allow the HNM for one direction to exist under the bivariate case, and serve as the most essential lemma for our identifiability theorem.

[5] Please refer to the arXiv version of this paper for supplementary materials.

Based on Eq. (9), we derive the negative log-likelihood of the marginal distribution $p(\mathbf{X})$ as the objective in our proposed formulation:

$$\mathcal{L}_{nll}(\mathbf{X}, A, B) = \sum_{m,n=1}^{M,N} \Big[\log\left(\sigma_n(\mathbf{X}(m), B_n)\sqrt{2\pi}\right) + \frac{\left(\mathbf{X}_n(m) - f_n(\mathbf{X}(m), A_n)\right)^2}{2\sigma_n^2(\mathbf{X}(m), B_n)} \Big] \tag{10}$$

The detailed derivation can be found in supplementary section 1.

Substituting the Eq. (7) and (8) into negative log-likelihood loss in Eq. (10), we obtain the training objective under proposed formulation w.r.t $W^{(1)}, W^{(2)}$, and $W^{(3)}$:

$$\begin{aligned} &\mathcal{L}_{nll}(\mathbf{X}, W^{(1)}, W^{(2)}, W^{(3)}) \\ &= \sum_{m,n=1}^{M,N} \Big[\log\sqrt{2\pi} + \log[\text{ReLU}\Big(W_n^{(3)} s\big(W_n^{(1)}\mathbf{X}^T(m)\big)\Big) + e^{W_{n0}^{(3)}}] \\ &\quad + \frac{\Big(\mathbf{X}_n(m) - W_n^{(2)} s\big(W_n^{(1)}\mathbf{X}^T(m)\big)\Big)^2}{2\big[\text{ReLU}\Big(W_n^{(3)} s\big(W_n^{(1)}\mathbf{X}^T(m)\big)\Big) + e^{W_{n0}^{(3)}}\big]^2} \Big] \end{aligned} \tag{11}$$

The DAG learning problem becomes the constrained continuous optimization that finds the optimal values $(W^{(1)})^*, (W^{(2)})^*, (W^{(3)})^*$ by minimizing $\mathcal{L}_{nll}(\mathbf{X}, W^{(1)}, W^{(2)}, W^{(3)})$ subject $h(W^{(1)}) = 0$.

Intuitively, by introducing and estimating conditional distribution variances $\boldsymbol{\sigma} = \{\sigma_n^2(\mathbf{X}(m), B_n)\}_{n,m=1}^{N,M}$ as functions of causes in HNM, our formulation allows the modeling of heteroscedasticity within the data noise. However, on the other hand, $\boldsymbol{\sigma}$ estimation inevitably increases modeling and optimization difficulties significantly, causing state-of-art global DAG learning methods like GraN-DAG++ (Lachapelle et al. 2019) to fail.

The difficulty of learning the causal DAG under the proposed formulation lies in effectively minimizing the negative log-likelihood loss over two sets of parameters A and B jointly while the interplay between optimization over A and B compromises the accuracy of each other. If the algorithm jointly learns A, B, the optimization process tends to minimize the negative log-likelihood loss by learning a set of B that significantly increases the estimated $\boldsymbol{\sigma}$. As a result, the algorithm can reach a stationary solution without enforcing the residual errors to be small. To solve such difficulties, we propose a DAG learning approach based on a two-phase algorithm, which estimates causal functions parameters A and $\boldsymbol{\sigma}$ estimation parameters B alternatively and iteratively.

4 Two-phase iterative learning algorithm

As we mentioned above, we introduce and model the parameters for conditional distribution variances $\boldsymbol{\sigma}$ into our model. To avoid the interplay between optimization over mean and variance parameters of conditional distributions, we propose to first estimate the variances $\boldsymbol{\sigma}$ and then estimate mean parameters under fixed variance. To provide mathematical justification for such an iterative learning approach, we introduce posterior distribution for variance q in Eq. (12). For simplicity, we denote $\sigma_n^2(\mathbf{X}(m), B_n)$ in Eq. (9) as $\sigma_n^2(m)$, and $\boldsymbol{\sigma}^2(m) := \{\sigma_n^2(m)\}_{n=1}^N$, $\boldsymbol{\sigma}^2 := \{\boldsymbol{\sigma}^2(m)\}_{m=1}^M$. Hence we can write the marginal log-likelihood of $\mathbf{X}$ as follows:

$$\log p(\mathbf{X}|A) \geq \int_{\boldsymbol{\sigma}^2} q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q) \log \frac{p(\mathbf{X}, \boldsymbol{\sigma}^2|A)}{q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q)} \, \mathrm{d}\boldsymbol{\sigma}^2 \tag{12}$$

We drop the entropy term $q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q) \log q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q)$, since we consider the Θ_q is independent of current parameters A. The objective is to maximize the lower bound of the marginal log-likelihood:

$$A^* = \arg\max_A \int_{\boldsymbol{\sigma}^2} q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q) \log p(\mathbf{X}, \boldsymbol{\sigma}^2|A) \, \mathrm{d}\boldsymbol{\sigma}^2 \tag{13}$$

We use t as the notation for the iteration index of our proposed algorithm. We chose Θ_q^t to be A^{t-1}, i.e., set Θ_q in the current iteration with A from the previous iteration, $q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q^t) = p(\boldsymbol{\sigma}^2|\mathbf{X}, A^{t-1})$. This selection of q has been proven to be a tight lower bound of p. To simplify the learning procedure, we obtain the optimal value of the $\boldsymbol{\sigma}^2$, denoted as $\hat{\boldsymbol{\sigma}}^2$, via maximizing the $p(\boldsymbol{\sigma}^2|\boldsymbol{X}, A^{t-1})$. Phase-I and Phase-II can be performed as follows.

$$\text{Phase-I}: \qquad \hat{\boldsymbol{\sigma}}^2 = \arg\max_{\boldsymbol{\sigma}^2} p(\boldsymbol{\sigma}^2|\mathbf{X}, A^{t-1}) \tag{14}$$

$$\text{Phase-II}: \qquad A^* = \arg\max_A \log p(\mathbf{X}, \hat{\boldsymbol{\sigma}}^2|A) \tag{15}$$

In Phase-I, to obtain the posterior distribution $p(\boldsymbol{\sigma}^2|\mathbf{X}, A^{t-1})$, we assume there exists a non-informative uniform prior $p(\boldsymbol{\sigma}^2)$[6]. Then the posterior distribution is proportional to the likelihood of the marginal distribution $p(\mathbf{X}|\boldsymbol{\sigma}^2, A^{t-1})$, i.e., $p(\boldsymbol{\sigma}^2|\mathbf{X}, A^{t-1}) \propto p(\mathbf{X}|\boldsymbol{\sigma}^2, A^{t-1})$. The optimal estimation of variances can be obtained by maximizing the likelihood of the marginal distribution $p(\mathbf{X}|\boldsymbol{\sigma}^2, A^{t-1})$, or minimizing its log-likelihood, i.e., the NLL loss in Eq. (10) with $A = A^{t-1}$. Given $\mathbf{X}$, the values of $\boldsymbol{\sigma}^2$ depend on the parameters in B that are not shared with A. The optimization in Eq. (14) can be simplified to set $W^{(1)} = (W^{(1)})^{t-1}, W^{(2)} = (W^{(2)})^{t-1}$ and optimize $W^{(3)}$ over $\mathcal{L}_{nll}$:

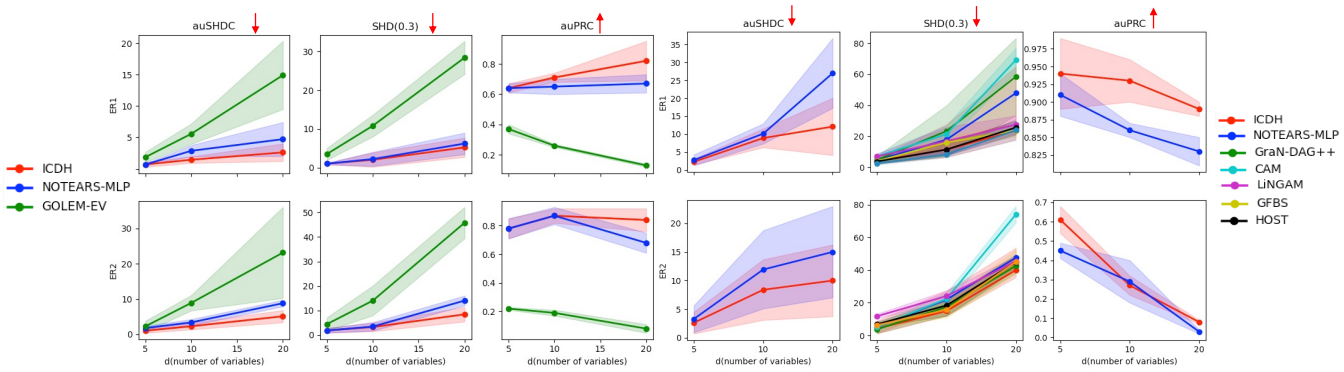
$$\begin{aligned} (W^{(3)})^* &= \arg\min_{W^{(3)}} \mathcal{L}_{nll}(\mathbf{X}, (W^{(1)})^{t-1}, (W^{(2)})^{t-1}, W^{(3)}) \\ \hat{\sigma}_n^2(m) &= \sigma_n(\mathbf{X}(m), (W^{(1)})^{t-1}, (W^{(3)})^*) \\ & n = 1, 2, \cdots, N, m = 1, 2, \cdots, M \end{aligned} \tag{16}$$

In Phase II, we directly maximize the likelihood given the optimal estimation of variances, or in practice minimize the NLL loss in Eq. (10) given $\boldsymbol{\sigma}^2 = \hat{\boldsymbol{\sigma}}^2$. The optimization in Eq. (15) can be simplified to optimize $W^{(1)}, W^{(2)}$ in A over $\mathcal{L}_{nll}$ with fixed values for variances and subject to the acyclicity constraint on $W^{(1)}$:

$$\begin{aligned} (W^{(1)})^*, (W^{(2)})^* = \arg\min \mathcal{L}_{nll}(\mathbf{X}, W^{(1)}, W^{(2)}, \hat{\boldsymbol{\sigma}}^2) \\ \text{subject to } h\big(W^{(1)}\big) = 0 \end{aligned} \tag{17}$$

We choose to only update $W^{(3)}$ in Phase-I to prevent poor empirical performance caused by the joint optimization over

[6]We choose a non-information prior for $p(\boldsymbol{\sigma}^2)$, which is the least restrictive so that we can simplify the objective into mere likelihood. Our formulation can also adapt to other types of the prior distribution.



(a) Results on nonlinear synthetic homoscedastic noise data

(b) Results on nonlinear synthetic heteroscedastic noise data

Figure 1: Comparison of corresponding baselines on nonlinear synthetic datasets: results (mean, standard error) on auSHDC, SHD, and auPRC. Our method is shown in the red curve.

two competing terms in our loss. If $W^{(1)}$ is jointly optimized in the Phase-I, we will obtain a degenerate solution with large reconstruction loss and larger unreasonable variances. To solve the constrained continuous optimization problem in Phase-II, we adopt a standard Lagrangian optimization process and force $W^{(1)}$ to satisfy the acyclicity constraint (Algorithm 3). The augmented Lagrangian optimization method is generally accepted as the better method, compared to the alternative penalty method (Ng et al. 2022). We choose ALM for a fair comparison since it has been employed by many state-of-art methods that tackle the same issue as our method. We outline the full procedure (Algorithm 1), Phase-I procedure (Algorithm 2), Phase-II procedure (Algorithm 3) in supplementary section 5.1.

Convergence analysis: Our proposed two-phase iterative learning approach can only guarantee to obtain a stationary solution, i.e., the gradients of parameters A and B w.r.t our training objective can achieve zeros after the algorithm converges. Please refer to supplementary in section 2.2 for details.

Complexity analysis: In Phase-II, the time complexity is $\mathcal{O}(N^3)$ w.r.t number of nodes N, which takes the same number of optimization iterations as other continuous methods with Augmented Lagrangian Method (ALM) (Zheng et al. 2020; Lachapelle et al. 2019). Phase I is relatively much cheaper in computation. The time complexity is $\mathcal{O}(mN^2)$ as one iteration of LBFGS with memory size m is employed. The total time complexity of our algorithm is $\mathcal{O}(kN^3)$ with k iterations of two phases ($k \leq 5$ in practice). Our proposed method has the same order of magnitude as the other baseline methods. Hence our method can handle the same amount of variables.

5 Experiment

We perform experiments on real data and synthetic data to demonstrate the effectiveness of our proposed method. We denoted our method as **I**dentifiable **C**ausal **D**iscovery under **H**eteroscedastic data(ICDH). For more details on synthetic data generation procedure and evaluation metrics, please refer to the supplementary sections 6 and 7.

Baselines. We compare our method against DAG learning methods using continuous optimization that also relaxes the strong assumptions of SEM: GOLEM-NV-L1 (Ng, Ghassami, and Zhang 2020), GOLEM-EV-L1 (Ng, Ghassami, and Zhang 2020), GraN-DAG (Lachapelle et al. 2019), GraN-DAG++ (Lachapelle et al. 2019); the methods also address the heteroscedastic noise issue but under combinatorial optimization framework: HEC (Xu et al. 2022) and CAREFL (Khemakhem et al. 2021), GFBS (Rajendran et al. 2021) and HOST (Duong and Nguyen 2023); popular baselines NOTEARS-MLP (Zheng et al. 2020), CAM (Peters et al. 2014), LiNGAM (Shimizu 2014), and GES (Chickering 2002). Xu et al. (2022); Khemakhem et al. (2021) aim to learn pairwise causal relations instead of global graph structures. Hence we can only show the comparison on the cause-effect pairs dataset.

It is worth noting that our method is not developed for heterogeneous data and scale-invariant data. Since the tasks and underlying assumptions in methods designed for heterogeneous data (Huang et al. 2020; Zhou, He, and Ni 2022) and scale-invariant data (Reisach, Seiler, and Weichwald 2021) deviate from ours, it is unfair to compare our method to those on synthetic data generated to solve our problems. The heteroscedastic noise may cause these methods to incorrectly estimate the marginal variance and identify the wrong causal order. For a comprehensive comparison, we experiment with CD-NOD and sortnregress methods on heteroscedastic noise data. Empirical results show that our method indeed performs better than these methods. Please refer to the supplementary material for details.

Moreover, we aim to develop a general and practical algorithm under HNM for static data with heteroscedastic noise. Hence we do not compare with methods that are either developed to learn temporal causal relations or employ complex noise distributions without explicitly modeling the variation of noise variances. Those are interesting research directions but are not relevant to this paper.

5.1 Empirical results on synthetic data

We generate synthetic data with different types of additive noises: homoscedastic noise with equal noise variances across variables and heteroscedastic noise. We also generated and experimented on homoscedastic noise with unequal noise variances across variables.

For each type of synthetic data, we compared different baselines based on the matchness between model formulations and data assumptions. The empirical results on homoscedastic equal noise data and heteroscedastic data are shown in Figure 1. Compared to the other SCM-based methods under a continuous optimization framework, empirical results indicate that our method can achieve comparable accuracy on homoscedastic noise data while outperforming baselines on heteroscedastic noise data. Compared to other types of methods, our method outperforms CAM, LiNGAM, and GFBS. Compared to GES and HOST, our method achieves comparable accuracy on data generated by sparse graphs and better performance on data generated on dense graphs. We also applied our method on larger dataset with 50 variables. On ER1 graphs, our ICDH method achieves SHD of 134.5 ± 23.4, outperform NOTEARS-MLP(144.1 ± 38.0), GraN-DAG++ (161.1 ± 30.8), and HOST(152.5 ± 24.8). The effectiveness of our method on dense graphs can be verified by empirical results on $ER3$ graphs. Please refer to the supplementary for the detailed numerical results.

5.2 Empirical results on real data

The empirical results on synthetic data, no matter homoscedastic or heteroscedastic, only indicate that the algorithms tend to perform well on the data that satisfies their model assumptions. These model assumptions are usually violated in real data or applications. Hence, a general formulation and an empirically effective learning approach are essential to solve real-world problems. We apply our method and the baseline methods on the two widely-studied real datasets: Sachs and cause-effect pairs.

Table 2: Comparison of SoTA methods on Sachs dataset.

Metrics	auSHDC↓	SHD(0.3)↓	SHD(optimal)↓	auPRC↑
NOTEARS-MLP	21.95	16	15	0.3427
GOLEM-EV	25.41	20	17	0.1697
GOLEM-NV	26.53	22	14	0.2524
GraN-DAG	-	-	**13**	-
GraN-DAG++	-	-	**13**	-
GFBS	-	-	17	-
HOST	-	-	**13**	-
Our Method	**19.27**	**15**	**13**	**0.4673**

Sachs Dataset: The results are summarized in Table 2. According to Table 2, we achieved a SHD of 16 for NOTEARS-MLP, which is close and lower than an SHD of 17 in their paper. GraN-DAG, GraN-DAG++, GFBS, and HOST employ a post-processing approach to find the optimal DAG with minimal SHD. We achieved SHDs of 13 for GraN-DAG, GraN-DAG++, and HOST, which are consistent with the SHDs from their original paper. For the GFBS method, we achieve the SHD of 17. Empirical results show that our proposed method can achieve comparable accuracy (SHD of 13) with SoTA methods and is more robust against thresholds.

Table 3: Comparison of SoTA methods on cause-effect pairs dataset: results on number of correct inferences of cause-effect relations and the weighted accuracy in general.

Methods	# of Correct inference↑	Weighted Accuracy ↑
NOTEARS-MLP	39/99	0.49
NOTEARS	36/99	0.47
GOLEM-EV	33/99	0.40
GOLEM-NV	33/99	0.40
ICDH(ours)	52/99	0.58
HEC	-	0.71 [7]
CAREFL	-	0.73 [8]

Cause-effect pairs dataset: Following the standard experiment procedures, we exclude the multivariate data sets and only experiment on the remaining 99 bivariate problems. The results are summarized in Table 3. Our method correctly inferred 52 out of 99 pairs of cause-effect relations while NOTEARS-MLP, NOTEARS, GOLEM-EV and GOLEM-NV correctly identified 39, 36, 33, and 33 pairs. However, our method achieve lower weighted accuracy compared to the reported result of HEC and CAREFL. Even though those methods make similar model assumptions as our method, they are developed specifically for bi-variate data by directly comparing two models $X \leftarrow Y$ and $X \rightarrow Y$ and selecting the one with a higher proposed objective value. Our whole DAG learning method requires continuous optimization, whereby may not find the global optimal objective. Hence they outperform all the whole DAG learning methods on the cause-effect pairs dataset. The empirical results in Tables 2-3 indicate that the real data is highly likely to have heteroscedastic variables with varying noise variances across samples. Our DAG learning method with a general model formulation and effective learning approach is more suitable for real-world applications.

6 Conclusion

In this paper, we introduce relaxed implementable sufficient conditions to provide the identifiability for a general class of multivariate SEM. We propose a novel formulation for the DAG learning problem guided by the conditions, which accounts for the noise variance variation across both variables and observations. Our formulation is identifiable and can generalize existing formulations of state-of-art methods. We then propose an effective two-phase iterative DAG learning approach to address the increasing training difficulties introduced by the general formulation. Empirical results show that our method achieves comparable accuracy on homoscedastic noise data while outperforming the SOTA methods on heteroscedastic noise data and real data, which indicates 1) the existing methods likely suffer when noise variances vary across observations, 2) our method has great potential for real-world applications.

[7] Reported results from (Xu et al. 2022)

[8] Reported results from (Khemakhem et al. 2021)

Acknowledgement: This work is supported in part by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network, and by the National Science Foundation award IIS 2236026.

References

Ahuja, K.; Caballero, E.; Zhang, D.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. *arXiv preprint arXiv:2106.06607*.

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

Chen, W.; Drton, M.; and Wang, Y. S. 2019. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4): 973–980.

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

Duong, B.; and Nguyen, T. 2023. Heteroscedastic Causal Structure Learning. *arXiv preprint arXiv:2307.07973*.

Gao, M.; Ding, Y.; and Aragam, B. 2020. A polynomial-time algorithm for learning nonparametric causal graphs. *arXiv preprint arXiv:2006.11970*.

Huang, B.; Zhang, K.; Zhang, J.; Ramsey, J. D.; Sanchez-Romero, R.; Glymour, C.; and Schölkopf, B. 2020. Causal Discovery from Heterogeneous/Nonstationary Data. *J. Mach. Learn. Res.*, 21(89): 1–53.

Immer, A.; Schultheiss, C.; Vogt, J. E.; Schölkopf, B.; Bühlmann, P.; and Marx, A. 2022. On the Identifiability and Estimation of Causal Location-Scale Noise Models. *arXiv preprint arXiv:2210.09054*.

Janzing, D.; and Schölkopf, B. 2018. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, 2245–2253. PMLR.

Javidian, M. A.; Pandey, O.; and Jamshidi, P. ???? Scalable Causal Domain Adaptation.

Kalainathan, D.; Goudet, O.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Sam: Structural agnostic model, causal discovery and penalized adversarial learning.

Khemakhem, I.; Monti, R.; Leech, R.; and Hyvarinen, A. 2021. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, 3520–3528. PMLR.

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

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

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.

Ng, I.; Lachapelle, S.; Ke, N. R.; Lacoste-Julien, S.; and Zhang, K. 2022. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, 8176–8198. PMLR.

Ott, S.; Imoto, S.; and Miyano, S. 2004. Finding optimal models for small gene networks. In *Pacific symposium on biocomputing*.

Park, G. 2020. Identifiability of Additive Noise Models Using Conditional Variances. *J. Mach. Learn. Res.*, 21(75): 1–34.

Peters, J.; Janzing, D.; and Scholkopf, B. 2011. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450.

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053.

Rajendran, G.; Kivva, B.; Gao, M.; and Aragam, B. 2021. Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families. *Advances in Neural Information Processing Systems*, 34: 18660–18672.

Reisach, A.; Seiler, C.; and Weichwald, S. 2021. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34: 27772–27784.

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

Schölkopf, B. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Sgouritsa, E.; Janzing, D.; Hennig, P.; and Schölkopf, B. 2015. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*, 847–855. PMLR.

Shen, Z.; Cui, P.; Kuang, K.; Li, B.; and Chen, P. 2018. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, 411–419.

Shimizu, S. 2014. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1): 65–98.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

Spirtes, P.; Meek, C.; and Richardson, T. 1995. Causal inference in the presence of latent variables and selection bias. In *UAI*.

Stojanov, P.; Li, Z.; Gong, M.; Cai, R.; Carbonell, J.; and Zhang, K. 2021. Domain Adaptation with Invariant Representation Learning: What Transformations to Learn? *Advances in Neural Information Processing Systems*, 34.

Wu, C. J. 1983. On the convergence properties of the EM algorithm. *The Annals of statistics*, 95–103.

Xu, S.; Marx, A.; Mian, O.; and Vreeken, J. 2022. Causal Inference with Heteroscedastic Noise Models.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG Structure Learning with Graph Neural Networks. *arXiv preprint arXiv:1904.10098*.

Yu, Y.; and Gao, T. 2020. DAGs with No Curl: Efficient DAG Structure Learning. *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Causal Discovery and Causality-Inspired Machine Learning*.

Zhang, J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896.

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

Zhang, K.; Zhang, J.; and Schölkopf, B. 2015. Distinguishing cause from effect based on exogeneity. *arXiv preprint arXiv:1504.05651*.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, 9472–9483.

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

Zhou, F.; He, K.; and Ni, Y. 2022. Causal Discovery with Heterogeneous Observational Data. *arXiv preprint arXiv:2201.12392*.

A The Derivation of the NLL Loss under the Proposed Formulation

In this section, we provide the detailed derivation of the NLL loss in Eq. (10) and Eq. (11) in the main paper.

$$
\begin{aligned}
&\mathcal{L}_{nll}(\boldsymbol{X}, A, B) \\
&= -\log p(\mathbf{X}) \\
&= -\log \prod_{m=1}^{M} p(\mathbf{X}(m)) \\
&= -\log \prod_{m=1}^{M} \prod_{n=1}^{N} p(\mathbf{X}_n(m)|\mathbf{X}_{\pi_n}(m)) \\
&= -\log \prod_{m=1}^{M} \prod_{n=1}^{N} \frac{1}{\sigma_n\big(\mathbf{X}(m), B_n\big)\sqrt{2\pi}} e^{-\frac{\left(\mathbf{x}_n(m) - f_n(\mathbf{X}(m), A_n)\right)^2}{2\sigma_n^2\left(\mathbf{x}(m), B_n\right)}} \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} \Big[\log\Big(\sigma_n\big(\mathbf{X}(m), B_n\big)\sqrt{2\pi}\Big) + \frac{\big(\mathbf{X}_n(m) - f_n(\mathbf{X}(m), A_n)\big)^2}{2\sigma_n^2\big(\mathbf{X}(m), B_n\big)} \Big] \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} \Big[\log\sqrt{2\pi} + \log[\mathrm{ReLU}\big(W_n^{(3)} s(W_n^{(1)}\mathbf{X}^T(m))\big) + e^{W_{n0}^{(3)}}] \\
&+ \frac{\Big(\mathbf{X}_n(m) - W_n^{(2)} s\big(W_n^{(1)}\mathbf{X}^T(m)\big)\Big)^2}{2\Big(\mathrm{ReLU}\big(W_n^{(3)} s(W_n^{(1)}\mathbf{X}^T(m))\big) + e^{W_{n0}^{(3)}}\Big)^2} \Big]
\end{aligned}
\tag{18}
$$

where $s(\cdot)$ is the sigmoid function. If we adopt the simplified notation, Eq. (18) can also be written as $\mathcal{L}_{nll}(\boldsymbol{X}, A, B) = \sum_{m=1}^{M}\sum_{n=1}^{N}[\log(\sigma_n(m)\sqrt{2\pi}) + \frac{(X_n(m) - f_n(\mathbf{X}(m), A_n))^2}{2\sigma_n^2(m)}]$. Hence we also denote the our loss function as $\mathcal{L}_{nll}(\boldsymbol{X}, A, \boldsymbol{\sigma}^2)$ in the following sections.

B The Property of Proposed Two-phase Algorithm

B.1 Marginal Log-likelihood

Our novel formulation based on the general form of SEM in **Definition 1** introduces the modeling and estimation of noise variances $\boldsymbol{\sigma}^2$, which inevitably increases the optimization difficulties. Not only there are additional variances $\boldsymbol{\sigma}^2$ to learn, but the variances $\boldsymbol{\sigma}^2$ can compromise the accuracy of structural parameters A during training. The interplay between A and $\boldsymbol{\sigma}^2$ can lead the algorithm to converge to unexpected stationary solutions. To alleviate such an issue, we employ a two-phase optimization procedure, where we treat the noise variances $\boldsymbol{\sigma}^2$ as the unknown variables that need to be estimated simultaneously. The detailed derivations are shown in Eq. (19).

$$
\begin{aligned}
\log p(\mathbf{X}|A) &= \log \int_{\boldsymbol{\sigma}^2} p(\mathbf{X}, \boldsymbol{\sigma}^2|A)\, \mathrm{d}\boldsymbol{\sigma}^2 \\
&= \log \int_{\boldsymbol{\sigma}^2} q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q) \frac{p(\mathbf{X}, \boldsymbol{\sigma}^2|A)}{q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q)}\, \mathrm{d}\boldsymbol{\sigma}^2 \\
&\geq \int_{\boldsymbol{\sigma}^2} q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q) \log \frac{p(\mathbf{X}, \boldsymbol{\sigma}^2|A)}{q(\boldsymbol{\sigma}^2|\mathbf{X}, \Theta_q)}\, \mathrm{d}\boldsymbol{\sigma}^2
\end{aligned}
\tag{19}
$$

B.2 Convergence Guarantee

Our algorithm can only guarantee achieving a solution with derivatives of the likelihood being arbitrarily close to zero, i.e., the solution is a stationary point. We can view our Phase-I step (Eq. (15) in the main paper) as an attempt to construct a function $Q(A|A^{t-1}) = \log p(\mathbf{X}|\hat{\boldsymbol{\sigma}}^2, A)$ by finding the optimal values $\hat{\boldsymbol{\sigma}}^2$. Then Phase-II step can be seen as choosing the A^t to be any value in the set of A, denoted as Ω, which maximizes $Q(A|A^{t-1})$. Assume the optimal value of A in t^{th} iteration is selected from $\mathcal{M}(A^{t-1})$, where $\mathcal{M}(\cdot)$ a point-to-set map such that

$$
Q(A'|A^{t-1}) \geq Q(A^{t-1}|A^{t-1}) \forall A' \in \mathcal{M}(A^{t-1})
$$

We define the log-likelihood in Eq. (19) as $L(A)$, i.e., $L(A) = \log p(\boldsymbol{X}|A) = -\mathcal{L}_{nll}(\mathbf{X}, A, \boldsymbol{\sigma}^2)$. According to the Theorem 1 in (Wu 1983), if 1) $\mathcal{M}$ is a closed point-to-set map in the complement of $\mathcal{F}$, which is a set of stationary points in the interior of Ω, 2) $L(A^t) > L(A^{t-1})$ for all $A^{t-1} \notin \mathcal{F}$, then the limit points of $\{A^t\}_{t=1}^{T}$ are stationary points of L and $L(A^t)$ converges monotonically to $L^* = L(A^*)$ for a stationary point A^*. Wu (1983) also gives a sufficient condition for the closedness of $\mathcal{M}$: $Q(\psi|\phi)$ is continuous in both ψ and ϕ. Such a condition is easily satisfied since our Q is continuous in A. For 2), proving $L(A^t) > L(A^{t-1})$ is equivalent to prove $\mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^{t-1}) > \mathcal{L}_{nll}(\mathbf{X}, A^t, \left(\boldsymbol{\sigma}^2\right)^t)$. $\left(\boldsymbol{\sigma}^2\right)^t$ is obtained by finding the variance parameters B^t through minimizing $\mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, B)$,

hence we have

$$\mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^t) \leq \mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^{t-1}) \quad (20)$$

A^t is obtained by minimizing $\mathcal{L}_{nll}(\mathbf{X}, A, \left(\boldsymbol{\sigma}^2\right)^t)$, and we have

$$\mathcal{L}_{nll}(\mathbf{X}, A^t, \left(\boldsymbol{\sigma}^2\right)^t) \leq \mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^t) \quad (21)$$

Combine Eq. (20) and Eq. (21), we have

$$\begin{aligned} \mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^{t-1}) &\geq \mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^t) \\ &\geq \mathcal{L}_{nll}(\mathbf{X}, A^t, \left(\boldsymbol{\sigma}^2\right)^t) \end{aligned} \quad (22)$$

Note that the equalities can not be satisfied simultaneously, otherwise the algorithm converges at iteration $t-1$ and $A^{t-1} \in \mathcal{F}$. Therefore,

$$\mathcal{L}_{nll}(\mathbf{X}, A^{t-1}, \left(\boldsymbol{\sigma}^2\right)^{t-1}) > \mathcal{L}_{nll}(\mathbf{X}, A^t, \left(\boldsymbol{\sigma}^2\right)^t) \quad (23)$$

Since our proposed algorithm satisfies both 1) and 2), it can converge monotonically to an optimal L^* with a stationary point A^*.

C Identifiable multivariate HNM

Theorem 3.2 in main paper provides the relaxing and implementable sufficient conditions for multivariate HNM defined in **Definition 3.1**. We provide a sketch for the proof of **Theorem 3.2** and separate the proof into two components: (1) For bivariate case on variables X_1 and X_2, identify the assumptions that must hold so the HNMs for causal relations $X_1 \leftarrow X_2$ and $X_1 \rightarrow X_2$ both satisfy the given distribution $p(X_1, X_2)$. **We then identify the sufficient conditions that violate the assumptions from (1). If those conditions are satisfied, then only the HNM for the true causal direction satisfies the data distribution. Then HNM is identifiable subject to those conditions for the bivariate cases.** (2) We follow the standard approach to prove the identifiability for multivariate cases using the identifiability theorem for bivariate cases.

First, we propose **Lemma C.1** to prove the identifiability of the HNM for bivariate cases subject to certain conditions. We aim to identify conditions that are easy to model and estimate in practice.

Lemma C.1. *Assume a random set with two variables $X = (X_1, X_2)$ follows the HNM described by Eq. (5) , with E_1, E_2 be the independent exogenous noise variables for X_1, X_2. E_1, E_2 follow Gaussian distributions. If functions f_j, σ_j linking cause to effect satisfy 1) f_j is nonlinear, and 2) σ_j is a piece-wise function, then the HNM is identifiable.*

Proof. For variables X_1, X_2, assume they follow the model

$$X_2 = f_2(X_1) + \sigma_2(X_1)E_2 \quad (24)$$

where E_2 is a standard Gaussian distribution, f_2, σ_2 are twice-differentiable scalar functions and $\sigma_2(X_1) > 0$. If a backward model exists, i,e, the data also follows the same model in the other direction,

$$X_1 = f_1(X_2) + \sigma_1(X_2)E_1 \quad (25)$$

where E_1 is a standard Gaussian distribution, f_1, σ_1 are twice-differentiable scalar functions and $\sigma_1(X_2) > 0$. The assumptions that must hold so the forward and backward models co-exist have been studied and identified by Khemakhem et al. (2021) and Immer et al. (2022). We employ theoretical results from **Theorem 2** in Khemakhem et al. (2021). If Eq. (24) and Eq. (25) co-exist, then one of the following scenarios must hold: (1) $(\sigma_2, f_2) = (\frac{1}{Q}, \frac{P}{Q})$ and $(\sigma_1, f_1) = (\frac{1}{Q'}, \frac{P'}{Q'})$ where Q, Q' are polynomials of degree two, $Q, Q' > 0$, P, P' are polynomials of degree two or less, and $p(X_1), p(X_2)$ are strictly log-mix-rational-log. (2) σ_1, σ_2 are constant, f_1, f_2 are linear and $p(X_1), p(X_2)$ are Gaussian densities.

By making f_j to be nonlinear, scenario (2) does not apply to our HNM. We then choose σ_j not to be in the format of $\frac{1}{Q}, Q$ are polynomials of degree two. Hence we let σ_j be a piece-wise function. For example, in our formulation, for causal direction $X_1 \rightarrow X_2$, we choose $\sigma_2(X_1) = \text{ReLU}\big(w_3 s(w_1 X_1)\big) + e^{w_{30}}$, where $e^{w_{30}}$ is to make sure $\sigma_2(X_1) > 0$, $s(\cdot)$ is sigmoid activation function. If conditions on f_j, σ_j are satisfied, then both scenarios do not hold for our HNM. There is no backward model for any distribution that satisfies Eq. (5) for bivariate cases. Hence, the model is identifiable. □

Compared to our identifiable conditions in **Lemma C.1**, conditions in Khemakhem et al. (2021) ensure scenarios (1) and (2) do not hold by choosing f_j to be nonlinear and invertible. However, the invertibility condition cannot be readily adapt to multivariate cases due to the different number of dimensions between inputs and outputs of function f_j s. Our piecewise σ conditions are more relxed and implementable on multivariate cases.

We then prove the identifiability of multivariate HNM using **Lemma C.1**. To prove that our HNM is identifiable for multivariate cases is to prove that a unique graph $\mathcal{G}$ can be identified subject to HNM. In the following proof, we employ the **Proposition 28**, **Lemma 35**, and **Lemma 36** from Peters et al. (2014). We show the theoretical results from those proposition and lemmas in the format of our HNM.

Assume that there exists another HNM with graph $\mathcal{G}'$ that $\mathcal{G} \neq \mathcal{G}'$. According to the **Propostion 28** in Peters et al. (2014), let $\mathcal{G}$ and $\mathcal{G}'$ be two different DAGs over a set of variables $\boldsymbol{X}$. Assume $p(\boldsymbol{X})$ is generated by our HNM and satisfies the Markov condition and causal minimality with respect to $\mathcal{G}$ and $\mathbf{G}'$. Then there are variables $L, Y \in \boldsymbol{X}$ such that for the set $\boldsymbol{Q} := \mathbf{PA}_Y^{\mathcal{G}} \backslash \{L\}$, $\boldsymbol{R} := \mathbf{PA}_L^{\mathcal{G}'} \backslash \{Y\}$ and $\boldsymbol{S} := \boldsymbol{Q} \cup \boldsymbol{R}$, we have: A) $L \rightarrow Y$ in $\mathcal{G}$ and $Y \rightarrow L$ in $\mathcal{G}'$. B) $\boldsymbol{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}} \backslash \{L\}$ and $\boldsymbol{S} \subseteq \mathbf{ND}_L^{\mathcal{G}'} \backslash \{Y\}$. $\mathbf{PA}_Y^{\mathcal{G}}$ is the set of parent variables of Y in graph $\mathcal{G}$. $\mathbf{ND}_Y^{\mathcal{G}}$ is the set of non-descendant variables of Y in graph $\mathcal{G}'$.

We consider $\boldsymbol{S} = \boldsymbol{s}$ with $p(\boldsymbol{s}) > 0$. Denote $L^* := L|\boldsymbol{S} = \boldsymbol{s}$ and $Y^* := Y|\boldsymbol{S} = \boldsymbol{s}$. **Lemma 36** in (Peters et al. 2014) states that if $p(\boldsymbol{X})$ is generated according to the SEM models in Eq. (26):

$$X_n = g_n(X_{\pi_n}, E_n), n = 1, 2, \cdots, N, X_n \in \boldsymbol{X} \quad (26)$$

with corresponding DAG $\mathcal{G}$, then for a variable $X_n \in \boldsymbol{X}$, if $\boldsymbol{K} \subseteq \mathbf{ND}_{X_n}^{\mathcal{G}}$ then $E_{X_n} \perp\!\!\!\perp \boldsymbol{K}$. Our HNM can be viewed one specific class of the SEM in Eq. (26). Hence, **Lemma 36** holds under our HNM and renders $E_Y \perp\!\!\!\perp (L, \boldsymbol{S})$ and $E_L \perp\!\!\!\perp (Y, \boldsymbol{S})$.

Lemma 35 from (Peters et al. 2014) indicates that if $E_Y \perp\!\!\!\perp (Y, \boldsymbol{Q}, \boldsymbol{R})$ then for all $\boldsymbol{q}, \boldsymbol{r}$ with $p(\boldsymbol{q}, \boldsymbol{r}) > 0$, $g(Y, \boldsymbol{Q}, E_Y)|_{\boldsymbol{Q}=\boldsymbol{q}, \boldsymbol{R}=\boldsymbol{r}} = g(Y|_{\boldsymbol{Q}=\boldsymbol{q}, \boldsymbol{R}=\boldsymbol{r}}, \boldsymbol{q}, E_Y)$. We apply **Lemma 35** and obtain that

$$g(L, \boldsymbol{Q}, E_Y)|_{\boldsymbol{S}=\boldsymbol{s}} = g(L|_{\boldsymbol{S}=\boldsymbol{s}}, \boldsymbol{q}, E_Y) = g(L^*, \boldsymbol{q}, E_Y) \quad (27)$$

$$g(Y, \boldsymbol{R}, E_L)|_{\boldsymbol{S}=\boldsymbol{s}} = g(Y|_{\boldsymbol{S}=\boldsymbol{s}}, \boldsymbol{r}, E_L) = g(Y^*, \boldsymbol{r}, E_L) \quad (28)$$

Hence according to our definition, we have,

$$Y^* = f_Y(\boldsymbol{q}, L^*) + \sigma_Y(\boldsymbol{q}, L^*)E_Y, E_Y \perp\!\!\!\perp L^* \text{ in } \mathcal{G} \quad (29)$$

$$L^* = f_L(\boldsymbol{r}, Y^*) + \sigma_X(\boldsymbol{r}, Y^*)E_L, E_L \perp\!\!\!\perp Y^* \text{ in } \mathcal{G}' \quad (30)$$

However, the co-existence of both Eq. (29) and Eq. (30) contradicts our identifiability theorem for the bivariate cases. Therefore, the assumeption that there exists another HNM with graph $\mathcal{G}'$ that $\mathcal{G} = \mathcal{G}'$ does not hold. Only one unique DAG $\mathcal{G}$ can be identified from $p(\boldsymbol{X})$.

D Generality of the proposed method

D.1 Comparison with SoA Methods using Reconstruction Loss

Many existing methods (Zheng et al. 2018, 2020; Yu et al. 2019) adopt the reconstruction loss as the optimization objective, usually based on the SEM with additive noise. The score of the DAG learning problem, i.e., $F(A, \mathbf{X})$ in Eq. (3) of the main paper can be calculated through Eq. (31).

$$\begin{aligned} F(A, \boldsymbol{X}) &= \frac{1}{2M}\|\boldsymbol{X} - f(\boldsymbol{X}, A)\|_F^2 \\ &= \frac{1}{2M}\sum_{m=1}^{M}\sum_{n=1}^{N}\big(X_n(m) - f_n(\boldsymbol{X}(m), A_n)\big)^2 \end{aligned} \quad (31)$$

Substituting $\sigma_n^2(m) = \sigma^2$ into our training objective in Eq. (18), we can obtain

$$\begin{aligned} &\mathcal{L}_{nll}(\boldsymbol{X}, A, \sigma^2) \\ &= MN\log(\sqrt{2\pi\sigma^2}) + \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\big(X_n(m) - f_n(\boldsymbol{X}(m), A_n)\big)^2}{2\sigma^2} \\ &= \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\big(X_n(m) - f_n(\boldsymbol{X}(m), A_n)\big)^2}{2\sigma^2} + \text{const} \end{aligned} \quad (32)$$

With constant variance σ^2, optimizing the NLL loss in Eq. (32) is equivalent to optimizing the reconstruction loss in Eq. (31).

D.2 Comparison with SoA Methods Using Likelihood Loss under SEM with Additive Noise.

GOLEM-NV (Ng, Ghassami, and Zhang 2020) and GraN-DAG (Lachapelle et al. 2019) relax the equal noise variance assumption across variables under SEM with additive noise. Substitute $\sigma_n^2(m) = \sigma_n^2$ into Eq. (18), we obtain $\mathcal{L}_{nll}(\boldsymbol{X}, A, \boldsymbol{\sigma}^2)$ as:

$$\begin{aligned} &\mathcal{L}_{nll}(\boldsymbol{X}, A, \boldsymbol{\sigma}^2) \\ &= M\sum_{n=1}^{N}\log(\sqrt{2\pi\sigma_n^2}) + \sum_{n=1}^{N}\frac{\sum_{m=1}^{M}\big(X_n(m) - f_n(\boldsymbol{X}(m), A_n)\big)^2}{2\sigma_n^2} \end{aligned} \quad (33)$$

Eq. (33) is equivalent to the loss in Lachapelle et al. (2019). $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2)$ are treated as parameters and estimated during training. However, if we choose causal function $f_n(\cdot)$ in Eq. (33) as a linear function, i.e., $f_n(\boldsymbol{X}(m), A_n) = \boldsymbol{X}(m)A_n$, then we can show that our derived loss is equivalent to the loss derived in Ng, Ghassami, and Zhang (2020). As shown in Ng, Ghassami, and Zhang (2020), when considering $\mathcal{L}_{nll}$ as a function of σ_n^2, its local extreme values occur when $\frac{\partial \mathcal{L}_{nll}}{\partial \sigma_n^2} = 0$, i.e.,

$$\hat{\sigma}_n^2 = \frac{\sum_{m=1}^{M}\Big[\big(X_n(m) - \boldsymbol{X}(m)A_n\big)^2\Big]}{M} \quad (34)$$

Substitute Eq. (34) into Eq. (33), we obtain $\mathcal{L}_{nll}$ as:

$$\begin{aligned} &\mathcal{L}_{nll} \\ &= \frac{MN}{2}\Big(1 + \log(2\pi) - \log(M)\Big) + \frac{M}{2}\sum_{n=1}^{N}\Big[\log\sum_{m=1}^{M}\big(X_n(m) - \boldsymbol{X}(m)A_n\big)^2\Big] \\ &= \frac{M}{2}\sum_{n=1}^{N}\Big[\log\sum_{m=1}^{M}\big(X_n(m) - \boldsymbol{X}(m)A_n\big)^2\Big] + \text{const} \end{aligned} \quad (35)$$

Eq. (35) is equivalent to the likelihood-based objective in Appendix C.1 of Ng, Ghassami, and Zhang (2020) under the assumption that A satisfies acyclicity constraint.

Hence, the losses that are derived under the additive noise SEM are merely special cases for the losses derived under more general SEM with affine noise.

E The Iterative DAG Learning Method

E.1 Procedures of the propose algorithm

We summarized the main procedure of our proposed two-phase iterative DAG learning algorithm in Algorithm 1, with Phase-I procedure in Algorithm 2 and Phase-II procedure in Algorithm 3.

F Dataset Description

F.1 Synthetic Data

To valid the effectiveness on various type of datasets, we apply our proposed algorithms on synthetic data, where various levels of noise heterosedacity are incorporated during generation process. We adopt the standard setup in (Zheng et al. 2020; Yu et al. 2019; Lachapelle et al. 2019). The ground-truth DAGs are generated from Erdo-Renyi (ER) with k expected edges, which we set as 1 and 2. We generate 10 graphs for each graph setting with different numbers of variables $d = 5, 10, 20, 50$. For each setting, we simulate 10 trials with $n = 1000$ data observations.

Algorithm 1: Main Procedure

Input: Data $\mathbf{X}$
Output: $(W^{(1)})^*, (W^{(2)})^*, (W^{(3)})^*$
Initial $(W^{(1)})^0, (W^{(2)})^0, (W^{(3)})^0$ with 0
$\left(\hat{\boldsymbol{\sigma}}\right)^0 \leftarrow \sigma\left(\mathbf{X}, (W^{(1)})^0, (W^{(3)})^0\right)$
$(W^{(1)})^1, (W^{(2)})^1 \leftarrow$ `Phase-II-Update`$(\mathbf{X}, (\hat{\boldsymbol{\sigma}})^0)$
$t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  {Phase-I step:}
  STATE $(W^{(3)})^t =$ `Phase-I-Update`$\left(\mathbf{X}, (W^{(1)})^{t-1}, (W^{(2)})^{t-1}\right)$
  $\left(\hat{\sigma}\right)^t \leftarrow \sigma\left(\mathbf{X}, (W^{(1)})^{t-1}, (W^{(3)})^t\right)$
  {Phase-II step:}
  $(W^{(1)})^t, (W^{(2)})^t =$ `Phase-II-Update`$\left(\mathbf{X}, (\hat{\boldsymbol{\sigma}})^t\right)$
until Converge
$(W^{(1)})^*, (W^{(2)})^*, (W^{(3)})^* \leftarrow (W^{(1)})^t, (W^{(2)})^t, (W^{(3)})^t$
Return $(W^{(1)})^*, (W^{(2)})^*, (W^{(3)})^*$

Algorithm 2: Phase-I-Update Procedure

Input: Data $\mathbf{X}$, $\hat{W}^{(1)}$, $\hat{W}^{(2)}$
Output: $(W^{(3)})^*$
Initial $(W^{(3)})^0$ with small values
$(W^{(3)})^* = \arg\min_{W^{(3)}} \mathcal{L}_{nll}(\mathbf{X}, \hat{W}^{(1)}, \hat{W}^{(2)}, W^{(3)})$
Return $(W^{(3)})^*$

Algorithm 3: Phase-II-Update Procedure

Input: Data $\mathbf{X}$; Noise variances $\hat{\boldsymbol{\sigma}}^2$
Output: $(W^{(1)})^*, (W^{(2)})^*$
Initial $(W^{(1)})^0, (W^{(2)})^0$ with small values.
$\alpha = 0, \rho = 1, t \leftarrow 0$
while $h\left((W^{(1)})^t\right) > \epsilon$ **do**
  while $\rho < \rho_{max}$ **do**

$$(W^{(1)})^c, (W^{(2)})^c = \arg\min_{W^{(1)}, W^{(2)}} \mathcal{L}_{nll}(\mathbf{X}, W^{(1)}, W^{(2)}, \hat{\boldsymbol{\sigma}}^2) + \frac{\rho}{2} h^2\left((W^{(1)})^c\right) + \alpha h\left((W^{(1)})^c\right) \tag{36}$$

    if $h\left((W^{(1)})^c\right) < c \cdot h\left((W^{(1)})^t\right)$ **then**
      $(W^{(1)})^{t+1}, (W^{(2)})^{t+1} \leftarrow (W^{(1)})^c, (W^{(2)})^c$
    else
      $\rho \leftarrow s \cdot \rho$
    end if
  end while
  $\alpha \leftarrow \alpha + \rho h\left((W^{(1)})^{t+1}\right)$
  $t \leftarrow t + 1$
end while
$(W^{(1)})^*, (W^{(2)})^* \leftarrow (W^{(1)})^t, (W^{(2)})^t$
Return $(W^{(1)})^*, (W^{(2)})^*$

Synthetic homoscedastic noise data. We first conduct experiments on nonlinear synthetic homoscedastic noise data. We consider two types of homoscedastic data. The simpler version assumes that the noise corresponding to different variables across data samples has equal variance, i.e. noises are homoscedastic w.r.t both variables and observations. The model formulation of Zheng et al. (2020) satisfies such data generation process and we employ the similar procedures to generate nonlinear data with Gaussian noise. We denote such type as **homo-EV** data. Given a randomly generated binary DAG $\mathcal{G}$, the observations are sampled from the SEMs in Eq. (37) following the topological order induced by $\mathcal{G}$:

$$X_n = f_n(X_{pa(n)}) + Z_n, n = 1, 2, \cdots, N \tag{37}$$

where we chose $f_n(\cdot)$ to be randomly initialized MLPs with one hidden layer of size 100 and sigmoid activation. Z_n **are standard Gaussian noises, i.e.,** $Z_n \sim \mathcal{N}(0, 1)$. The slighter complex version, denoted as **homo-NV** data, allows the noises for different variables to have non-equal variances yet the noise variances across observations remain to be the same, i.e., $Z_n \sim \mathcal{N}(0, \sigma_n^2), n = 1, 2, \cdots, N$. We obtain the variances by sampling from a uniform distribution, i.e., $\sigma_n^2 \sim U[0.5, 2]$. We employ similar data generation process with Lachapelle et al. (2019) since its formulation fits assumptions.

Synthetic heteroscedastic noise data. We then evaluate our proposed algorithm on nonlinear synthetic heteroscedastic noise data. For heteroscedastic noise data, the noise variances vary across both variables and observations. Hence, heteroscedastic noise data is more challenging to accurately recover the DAG from the given observations. Given a random directed acyclic graph $\mathcal{G}$ with binary entries, we generate observations from the SEMs in Eq. (38) following the topological order induced by $\mathcal{G}$:

$$X_n = f_n(X_{pa(n)}) + e^{g_n(X_{pa(n)})} Z_n, n = 1, 2, \cdots, N \tag{38}$$

$f_n(\cdot)$ and $g_n(\cdot)$ are chosen to be randomly initialized MLPs with one hidden layer of size 100 and sigmoid activation. During the data generation process, we choose the variance function to be a global estimator, i.e., $\sigma_n = e^{g_n(\boldsymbol{X}_{pa(n)})}$ in order to test our piece-wise variance function's ability in recovering accurate variances. Z_n are standard Gaussian noises, $Z_n \sim \mathcal{N}(0, 1)$. We denote the data generated through above process as **hetero data**.

F.2 Real Data.

To demonstrate the effectiveness on real data, we test the proposed method on two widely-studied real benchmark datasets: Sachs dataset (Sachs et al. 2005) and cause-effect pairs dataset (Sgouritsa et al. 2015). The Sachs dataset consists of 7466 continuous measurements of expression levels of proteins and phospholipids in human immune system cells of 11 types. It is considered to have a census ground-truth causal network. The cause-effect pairs dataset provides 99 sets of data with given cause-effect relations between variables.

G Experiment Setting

Evaluation metrics. We employ 3 evaluation metrics to evaluate the accuracy of DAG learning: SHD, auSHDC, auPRC.

SHD: SHD is the most widely used evaluation metrics to evaluate the accuracy of a learned graph. However, a heuristic threshold approach needs to be performed in order to infer a DAG $\mathcal{G}$ from the weighted adjacency matrix. We report two SHDs in this paper: the SHD with threshold as 0.3, which is also chosen by the majority of the existing methods, and the minimum SHD obtained by using thresholds within the chosen range.

We also choose evaluation metrics that are less susceptible to thresholding.

auSHDC: To reduce the effect of thresholding on SHD, we choose a reasonable range for thresholds, estimate the SHD value of the graph thresholded with different thresholds, and plot the curve of SHDs versus thresholds. We employ the area under the SHD curve as a measurement of graph accuracy. A small auSHDC value indicates that the applied algorithm performs well and robust regardless of the thresholds. Since the synthetic graph parameters are from $U([-2.0, 0.5] \cup [0.5, 2.0])$, we believe $[0.2, 0.75]$ is a reasonable range for synthetic datasets. We adjust the range to be $[0.25, 0.75]$ for large models based on the empirical results.

auPRC: auPRC does not require to choose a constant value as threshold. The precision-recall curve (PRC) can be plotted from the learned weighted adjacency matrix. The accuracy performance can be reflected by the area under the precision-recall curve (auPRC). The graph with a larger auPRC values has a higher accuracy.

G.1 Implementation Details

We implemented the algorithm following the pseudo-code outlined in Algorithm 1, 2, and 3 in the main paper. We choose the LBFGS optimizer from the `scipy` library. For hyper-parameters in Algorithm 3, we set $\epsilon = 10^{-8}, c = 0.25, s = 10$ as suggested in the Zheng et al. (2018) where the augmented Lagrangian process for DAG learning is first introduced. We set the number of hidden neurons as $m_1 = 10$ for all the baselines. We conducted all experiments on a workstation with a 3.1 GHz CPU.

H Detailed empirical results

H.1 Nonlinear Synthetic Data

We show the detailed empirical results for homo-ev, homo-nv, and hetero data in Table 4, 5, and 6 respectively. We compared to baselines on different types of data depending on the matchness of their underlying model assumptions. However, since the GOLEM-EV and GOLEM-NV are implemented for data with linear relations, hence the results on nonlinear data are worse than other nonlinear DAG learning methods. In conclusion, we expect, and observed from three tables that our proposed method can achieve comparable results on data with equal noise variances across observations. Our proposed method outperforms the baselines on heteroscedastic data whereby the noise variances also vary with different values of causes.

We also performed experiments on larger datasets with $N = 50$ variables. Based on the data generation process that we elaborate above, a significant degree of noise has been embedded into the data, causing compromised performance on both our method and baseline methods. However, we will probably never expect such amout of data noise in real-world application.

I Limitations

In this paper, we propose a novel DAG learning formulation based on a general SEM which allows the modeling of the variation of noise variances across both variables and observations. To solve the increasing difficulties in optimization, we propose a two-phase iterative learning algorithm. However, there are two main limitations to the proposed algorithm. First, the proposed algorithm inevitably inherits the typical optimization difficulties for iterative optimization algorithms. The proposed iterative DAG learning algorithm only guarantees to converge to a stationary solution. Hence good initialization is crucial for the algorithm to achieve satisfactory performance. Another limitation is that our formulation has to satisfy the definition of HNM in **Definition 1** and is identifiable only when sufficient conditions in **Theorem 3.2** are satisfied.

		NOTEARS-MLP			GOLEM-EV			Our Method		
graph	d	auSHDC	SHD	auPRC	auSHDC	SHD	auPRC	auSHDC	SHD	auPRC
ER1	5	0.68 ± 0.15	1.0 ± 0.20	0.64 ± 0.03	1.85 ± 0.82	3.5 ± 1.55	0.38 ± 0.03	0.69 ± 0.39	1.0 ± 0.20	0.64 ± 0.03
	10	1.86 ± 0.90	2.2 ± 1.81	0.70 ± 0.05	5.56 ± 1.61	10.8 ± 2.81	0.27 ± 0.01	1.44 ± 0.52	2.0 ± 1.83	0.71 ± 0.02
	20	3.22 ± 2.70	6.2 ± 2.77	0.78 ± 0.06	14.92 ± 5.42	28.4 ± 4.27	0.13 ± 0.01	2.61 ± 1.39	5.2 ± 2.49	0.82 ± 0.13
	50	-	24.5 ± 6.20	-	-	50.6 ± 8.4	-	-	22.5 ± 5.5	-
		NOTEARS-MLP			GOLEM-EV			Our Method		
graph	d	auSHDC	SHD	auPRC	auSHDC	SHD	auPRC	auSHDC	SHD	auPRC
ER2	5	1.04 ± 0.61	1.80 ± 0.91	0.78 ± 0.07	2.28 ± 1.54	4.40 ± 2.72	0.2234 ± 0.01	1.03 ± 0.60	1.8 ± 0.98	0.78 ± 0.07
	10	2.35 ± 0.89	3.4 ± 1.74	0.87 ± 0.06	8.91 ± 2.17	17.2 ± 6.07	0.19 ± 0.02	2.31 ± 0.92	3.2 ± 1.72	0.87 ± 0.05
	20	7.98 ± 0.93	14.0 ± 2.00	0.68 ± 0.07	23.17 ± 12.88	45.8 ± 6.39	0.08 ± 0.03	5.13 ± 1.74	8.4 ± 3.98	0.84 ± 0.08
	50	-	39.0 ± 9.7	-	-	100.6 ± 8.4	-	-	30.9 ± 10.3	-

Table 4: Comparison of all baseline algorithms on nonlinear synthetic homoscedastic noise datasets with equal variances across variables and observations (homo-EV): results (mean $\pm$ standard deviation over 10 trails) on auSHDC, SHD, and auPRC.

		auSHDC			SHD			auPRC		
graph	methods	$d5$	$d10$	$d20$	$d5$	$d10$	$d20$	$d5$	$d10$	$d20$
ER1	Ours	0.41 ± 0.37	0.49 ± 0.26	2.53 ± 1.19	0.4 ± 0.80	0.6 ± 0.49	4.2 ± 1.94	0.71 ± 0.11	0.89 ± 0.02	0.79 ± 0.09
	NOTEARS-MLP	0.40 ± 0.38	0.53 ± 0.31	2.60 ± 1.19	0.4 ± 0.80	0.4 ± 0.49	4.2 ± 1.94	0.72 ± 0.10	0.89 ± 0.02	0.80 ± 0.08
	GOLEM-NV	2.14 ± 1.57	3.95 ± 2.81	10.25 ± 2.21	3.20 ± 3.71	5.60 ± 5.82	15.00 ± 5.48	0.85 ± 0.13	0.88 ± 0.10	0.92 ± 0.04
	GraN-DAG	-	-	-	2.4 ± 1.51	3.6 ± 1.52	6.2 ± 2.77	-	-	-
		auSHDC			SHD			auPRC		
graph	methods	$d5$	$d10$	$d20$	$d5$	$d10$	$d20$	$d5$	$d10$	$d20$
ER2	Ours	0.55 ± 0.98	2.83 ± 1.01	5.39 ± 2.31	1.0 ± 2.00	4.0 ± 2.61	8.6 ± 4.00	0.84 ± 0.13	0.78 ± 0.13	0.83 ± 0.08
	NOTEARS-MLP	0.55 ± 0.98	2.87 ± 0.10	5.63 ± 2.26	1.0 ± 2.00	4.0 ± 2.61	9.6 ± 4.30	0.84 ± 0.13	0.78 ± 0.13	0.82 ± 0.07
	GOLEM-NV	5.43 ± 1.12	20.67 ± 8.74	76.64 ± 18.72	7.20 ± 2.93	36.80 ± 19.33	149.00 ± 45.55	0.80 ± 0.07	0.62 ± 0.19	0.43 ± 0.14
	GraN-DAG	-	-	-	2.2 ± 2.95	9.4 ± 4.04	18.6 ± 7.73	-	-	-

Table 5: Comparison of all baseline algorithms on nonlinear synthetic homoscedastic noise datasets with equal variances across variables (homo-NV): results (mean $\pm$ standard deviation over 10 trails) on auSHDC, SHD, and auPRC.

		auSHDC				SHD				auPRC			
graph	methods	$d5$	$d10$	$d20$	$d50$	$d5$	$d10$	$d20$	$d50$	$d5$	$d10$	$d20$	$d50$
ER1	NOTEARS-MLP	2.75 ± 1.51	10.13 ± 2.79	27.02 ± 9.74	-	4.6 ± 1.51	20.6 ± 4.77	54.4 ± 27.00	144.1 ± 38.0	0.25 ± 0.03	0.25 ± 0.01	0.17 ± 0.02	-
	GOLEM-NV	2.29 ± 1.57	13.14 ± 3.25	15.23 ± 10.87	-	4.2 ± 1.99	20.6 ± 16.4	54.40 ± 24.20	-	0.37 ± 0.08	0.21 ± 0.11	0.83 ± 0.06	-
	GraN-DAG	-	-	-	-	6.2 ± 1.92	27.4 ± 7.99	86.0 ± 44.29	-	-	-	-	-
	GraN-DAG++	-	-	-	-	4.8 ± 2.28	22.2 ± 16.5	58.4 ± 24.79	161.1 ± 10.80	-	-	-	-
	CAM	-	-	-	-	7.1 ± 1.70	21.3 ± 5.48	69.2 ± 7.64	-	-	-	-	-
	LiNGAM	-	-	-	-	7.0 ± 1.61	16.8 ± 9.04	27.5 ± 5.70	-	-	-	-	-
	GES	-	-	-	-	4.0 ± 1.95	15.9 ± 3.59	23.7 ± 4.03	-	-	-	-	-
	GFBS	-	-	-	-	$\mathbf{3.7 \pm 1.77}$	11.6 ± 3.92	25.6 ± 4.58	-	-	-	-	-
	Ours	2.28 ± 0.84	8.91 ± 2.61	12.10 ± 7.98	-	4.0 ± 1.58	$\mathbf{11.4 \pm 4.93}$	$\mathbf{23.6 \pm 9.34}$	$\mathbf{134.5 \pm 23.40}$	0.39 ± 0.05	0.20 ± 0.03	0.20 ± 0.01	-
		auSHDC				SHD				auPRC			
graph	methods	$d5$	$d10$	$d20$	$d50$	$d5$	$d10$	$d20$	$d50$	$d5$	$d10$	$d20$	$d50$
ER2	NOTEARS-MLP	3.29 ± 2.36	11.93 ± 6.83	15.20 ± 7.97	-	5.20 ± 3.71	22.0 ± 5.30	47.60 ± 5.90	111.1 ± 11.80	0.61 ± 0.04	0.27 ± 0.11	0.08 ± 0.01	-
	GOLEM-NV	3.90 ± 2.30	20.05 ± 10.17	55.88 ± 38.91	-	6.20 ± 2.10	25.60 ± 5.94	67.60 ± 7.73	-	0.60 ± 0.07	0.47 ± 0.15	0.13 ± 0.16	-
	GraN-DAG++	-	-	-	-	$\mathbf{3.8 \pm 2.44}$	17.4 ± 5.15	42.6 ± 4.31	-	-	-	-	-
	GraN-DAG	-	-	-	-	$\mathbf{3.8 \pm 2.83}$	19.6 ± 4.65	77.60 ± 10.20	123.1 ± 9.60	-	-	-	-
	CAM	-	-	-	-	5.2 ± 2.71	22.4 ± 3.93	74.1 ± 7.60	-	-	-	-	-
	LiNGAM	-	-	-	-	11.8 ± 1.47	24.1 ± 3.36	45.5 ± 2.96	-	-	-	-	-
	GES	-	-	-	-	10.2 ± 2.04	21.5 ± 3.64	45.8 ± 4.98	-	-	-	-	-
	GFBS	-	-	-	-	6.9 ± 1.45	18.4 ± 3.78	44.9 ± 4.63	-	-	-	-	-
	Ours	2.63 ± 1.89	8.39 ± 5.28	10.10 ± 6.25	-	4.2 ± 3.06	$\mathbf{15.0 \pm 3.05}$	$\mathbf{40.0 \pm 4.40}$	$\mathbf{102.0 \pm 30.80}$	0.61 ± 0.07	0.21 ± 0.05	0.08 ± 0.01	-

Table 6: Comparison of all baseline algorithms on nonlinear synthetic heteroscedastic noise datasets: results (mean $\pm$ standard deviation over 10 trails) on auSHDC, SHD, and auPRC.