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.
yinn2@rpi.edu, tgao@us.ibm.com, yuy214@lehigh.edu, jiq@rpi.edu

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.

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 2021; 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.

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.

Background and Formulation

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.

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

Table 1: Summary of SEMs and algorithms for SoTA methods. "Homo" represents homoscedastic noise, and "Hetero" represents heteroscedastic noise. GFBS employs multiple linear and nonlinear SEMs, each with varying identifiability.

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.

Problem Statement

We first introduce one of the general SEM formulations in **Definition 1** that modulate the noise variance with cause variables. The SEMs we consider in **Definition 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.

[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.

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

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 2**. We can theoretically prove that the HNM is identifiable if those sufficient conditions hold.

Theorem 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 the 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.

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 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

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

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.

Optimization Objective and Difficulties

The goal is to estimate a DAG $\mathcal{G}$, given M observations of X, i.e., input matrix $\mathbf{X} = \{\mathbf{X}(m)\}_{m=1}^{M}$.

[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. https://arxiv.org/abs/2312.12844

$\mathbf{X}(m) \in \mathbb{R}^{1\times N}$ 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:

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

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

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

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 $W^{(3)}$:

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

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.

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.

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

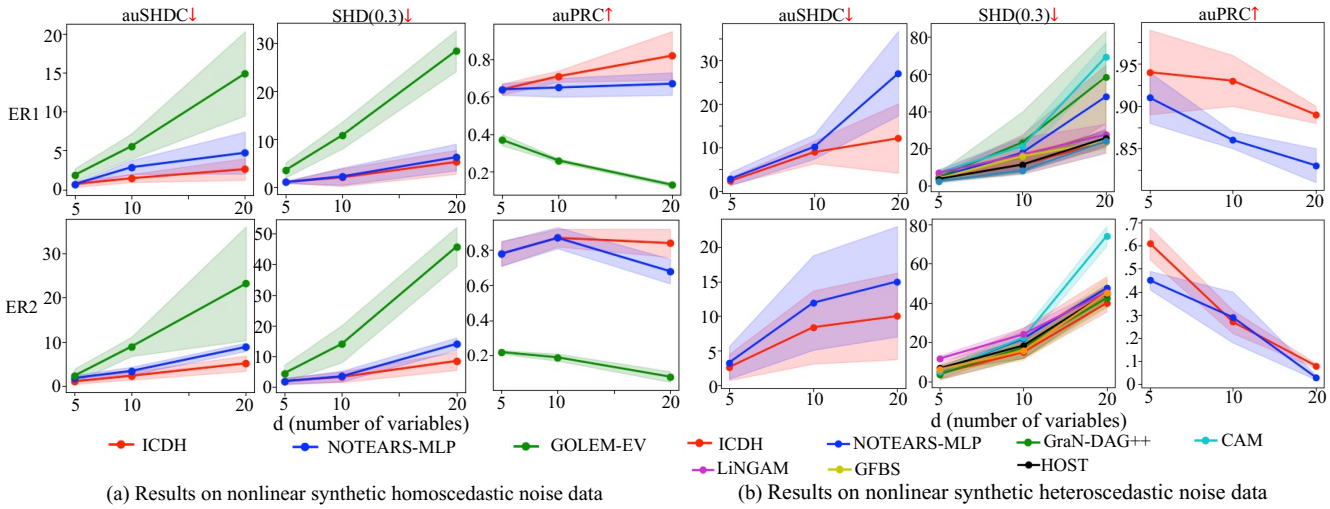
$$\text{Phase-II}: \qquad A^* = \arg\max_A \log p(\mathbf{X}, \hat{\boldsymbol{\sigma}}^2|A) \tag{15}$$

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 $\mathcal{L}_{nll}$:

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

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

[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 SoTA baselines on synthetic data: results(mean, standard error) on auSHDC, SHD, and auPRC. We only report auSHDC and auPRC on baselines directly return weighted adjacent matrices. Our method is shown in the red curve.

over $\mathcal{L}_{nll}$ with fixed values for variances and subject to the acyclicity constraint on $W^{(1)}$:

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

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

Convergence Analysis: Our proposed two-phase iterative learning approach can only guarantee 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 the supplementary Section B.2.

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, and can handle the same amount of variables.

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). Please refer to the supplementary Sections F, G and H for details on synthetic data generation, evaluation metrics, and numerical results.

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, thus are only compared on cause-effect pairs dataset.

Our method is not designed for heterogeneous and scale-invariant data. Tasks and assumptions in methods for heterogeneous data (Huang et al. 2020; Zhou, He, and Ni 2022) and scale-invariant data (Reisach, Seiler, and Weichwald 2021) differ from ours, making comparisons unfair on synthetic data tailored to our problems. Heteroscedastic noise may lead these methods to misestimate marginal variance and identify the wrong causal order. For a thorough comparison, we experiment with CD-NOD and sortnregress on heteroscedastic noise data (Supplementary Section H), demon-

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

Table 2: Comparison of SoTA methods on Sachs dataset.

Methods	Accuracy ↑	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]

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

strating our method's superior performance. Our focus is on developing a general algorithm under heterogeneous noise models for static data. Thus, we refrain from comparing with methods for temporal causal relations or those using complex noise distributions without explicitly modeling noise variance variation, as they are not relevant to this paper.

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.

We compared different baselines for each synthetic data type, considering the match between model formulations and data assumptions. Results on homoscedastic equal noise and heteroscedastic data are shown in Figure 1. In a continuous optimization framework, our method achieves comparable accuracy to other SCM-based methods on homoscedastic noise data and outperforms baselines on heteroscedastic noise data. Compared to CAM, LiNGAM, and GFBS, our method excels. Against GES and HOST, it achieves comparable accuracy on sparse graph-generated data and better performance on dense graph-generated data. The effectiveness of our method is also demonstrated on datasets with a larger number of variables ($d = 50$) and denser graphs (ER3). For more details, please refer to supplementary Section H.

Empirical Results on Real Data

Empirical results on synthetic data, whether homoscedastic or heteroscedastic, indicate algorithms perform well when data aligns with their model assumptions. However, these assumptions are often violated in real-world data. Therefore, a general formulation and an effective learning approach are essential. We apply our method and baselines to two real datasets: Sachs and cause-effect pairs.

Sachs Dataset: The results are summarized in Table 2. Our SHD of 16 for NOTEARS-MLP closely aligns with and is lower than the SHD of 17 reported in their paper. GraN-DAG, GraN-DAG++, GFBS, and HOST use post-processing to find the optimal DAG with minimal SHD. We achieve SHDs of 13 for GraN-DAG, GraN-DAG++, and HOST, consistent with their original papers. For the GFBS method, we achieve an SHD of 17. Empirical results demonstrate that our proposed method attains comparable accuracy (SHD of 13) with state-of-the-art methods and is robust against thresholds.

Cause-Effect Pairs Dataset: Following standard experimental procedures, we focus on the 99 remaining bivariate problems, as summarized in Table 3. Our method correctly infers 52 out of 99 cause-effect pairs, outperforming all the other whole DAG learning methods: NOTEARS-MLP, NOTEARS, GOLEM-EV, and GOLEM-NV, which correctly identify 39, 36, 33, and 33 pairs, respectively. Our method achieves a lower weighted accuracy compared to HEC and CAREFL. Despite similar model assumptions, these methods are tailored for bivariate data, directly comparing models $X \leftarrow Y$ and $X \rightarrow Y$ to select the one with a higher proposed objective value. Our whole DAG learning method, relying on continuous optimization, may not find the global optimal objective. Furthermore, empirical results in Tables 2-3 suggest real data likely involves heteroscedastic variables with varying noise variances across samples. Our DAG learning method, with a general model formulation and effective learning approach, proves more suitable for real-world data applications.

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)

Acknowledgements

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. 2021. Scalable Causal Domain Adaptation. *arXiv preprint arXiv:2103.00139*.

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.

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

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.

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*.