Journal of Machine Learning Research 21 (2020) 1-34 Submitted 8/19; Revised 4/20; Published 4/20

# Identifiability of Additive Noise Models Using Conditional Variances

**Gunwoong Park** GW.PARK23@GMAIL.COM
*Department of Statistics*
*University of Seoul*
*Seoul, 02504, South Korea*

**Editor:** Joris Mooij

## Abstract

This paper considers a new identifiability condition for additive noise models (ANMs) in which each variable is determined by an arbitrary Borel measurable function of its parents plus an independent error. It has been shown that ANMs are fully recoverable under some identifiability conditions, such as when all error variances are equal. However, this identifiable condition could be restrictive, and hence, this paper focuses on a relaxed identifiability condition that involves not only error variances, but also the influence of parents. This new class of identifiable ANMs does not put any constraints on the form of dependencies, or distributions of errors, and allows different error variances. It further provides a statistically consistent and computationally feasible structure learning algorithm for the identifiable ANMs based on the new identifiability condition. The proposed algorithm assumes that all relevant variables are observed, while it does not assume faithfulness or a sparse graph. Demonstrated through extensive simulated and real multivariate data is that the proposed algorithm successfully recovers directed acyclic graphs.

**Keywords:** Bayesian Network, Causal Inference, Directed Acyclic Graph, Identifiability, Structural Equation Modeling, Structure Learning

## 1. Introduction

Directed acyclic graphical (DAG) models, or Bayesian networks, are widely used to represent conditional independence and causal relations among random variables in many fields, such as meteorology, epidemiology, finance, genetics, neuroscience, sports science, and many others (Friedman et al. 2000; Peters and Bühlmann 2014; Sachs et al. 2005; Park and Raskutti 2018). However, learning directed graphical models is a notoriously difficult problem when interventional experiments are very expensive or impossible due to the identifiability issue and the super-exponentially growing size of the space of directed acyclic graphs in the number of nodes.

A number of prior works have tackled the non-identifiability problem for different classes of joint distributions by placing further restrictions. As a result, Spirtes et al. (2000), Chickering (2003), Tsamardinos and Aliferis (2003), Zhang and Spirtes (2016), Raskutti and Uhler (2018), and many other works show that DAG models are recoverable up to the Markov equivalence class under faithfulness or related assumptions. However, the true graph may not be uniquely determined, since most Markov equivalence classes contain more than one graph.

Hence, many recent works have attempted to find fully identifiable classes of DAG models by placing a different type of restrictions on the distributions. For example, Shimizu et al. (2006); Zhang and Hyvärinen (2009b) show that linear non-Gaussian additive noise models (ANMs) can be identifiable where each variable is determined by a linear function of its parents plus an independent error term. More precisely, the models are identifiable if one of its parents or error term belongs to a set of some non-Gaussian distributions. Zhang and Hyvärinen (2009a); Hoyer et al. (2009); Mooij et al. (2009); Peters et al. (2012) relax the assumption of linearity, and prove the identifiability of nonlinear ANMs where each variable is determined by a non-linear function of its parents and an error term. Park and Raskutti (2015, 2018); Park and Park (2019a,b) prove the identifiability of DAG models where a higher order moment of the conditional distribution of each node given its parents is a non-concave function of the mean; and Peters and Bühlmann (2014) prove that (Gaussian) linear structural equation models (SEMs) with equal or known error variances are identifiable. More recently, Ghoshal and Honorio (2018); Park and Kim (2020) show that Gaussian linear SEMs with unknown heterogeneous error variances can be identifiable. Mooij et al. (2016); Eberhardt (2017); Glymour et al. (2019) elegantly summarize the ideas of all these approaches using the notion of complexity or uncertainty. The uncertainty level of the conditional distribution of a node given its parents is, in general, lower than the conditional distribution given a strict subset of its parents, while all the approaches exploit the different uncertainty measures for the marginal and conditional probability distributions.

This paper proves the identifiability of *additive noise models* (ANMs) with any form of relationship between variables and unknown heterogeneous error variances by exploiting marginal and conditional variances. Our approach is a generalization of the identifiability result for linear SEMs in Peters and Bühlmann (2014); Loh and Bühlmann (2014); Ghoshal and Honorio (2017, 2018); Chen et al. (2019) where a (conditional) variance is used for the uncertainty measure of a (conditional) distribution. We provide a detailed comparison of our new condition to the previous identifiability conditions of linear SEMs in Section 3. However, the emphasis of our approach is not restricted to linear SEMs. Hence, our identifiable class of DAG models includes linear and non-linear SEMs with unknown heterogeneous error variances.

Further developed is a DAG structure learning algorithm, called Uncertainty Scoring (US), for learning the new identifiable ANMs based on the proposed identifiability condition. A notable point is that our identifiability condition enables the US algorithm to learn a DAG in two-steps. In the first step, the US algorithm estimates the ordering component wisely, either from the beginning or the end, using conditional variances. In the second step, the algorithm estimates the directed edges using conditional independence relationships. By decoupling the ordering estimation or parents search, the US algorithm gains significant computational improvements like many existing scalable DAG learning algorithms (e.g., Shimizu et al. 2011; Loh and Bühlmann 2014; Park and Raskutti 2018; Ghoshal and Honorio 2017; Park and Park 2019a; Wang and Drton 2020). Also provided is statistical consistency of our US algorithm in learning Gaussian linear SEMs.

The US algorithm is compared, through simulation studies and real multivariate data, against state-of-the-art greedy equivalence search (GES) (Chickering, 2003), greedy DAG search (GDS) (Peters and Bühlmann, 2014), linear structural equation model learning (LISTEN) (Ghoshal and Honorio, 2018), and linear non-Gaussian acyclic models (LINGAM) (Shimizu et al., 2006) algorithms in terms of the accuracy of recovering a graph structure. The simulation study first considers the various $p$-node Gaussian linear SEMs, with the number of nodes $p \in \{20, 200\}$, with the maximum number of parents $d \in \{2, 4\}$, and with error variances both homogeneous and heterogeneous. Also considered are non-Gaussian linear SEMs and Gaussian non-linear SEMs, where each variable is determined by a polynomial function of its parents plus a possibly non-Gaussian additive error.

The remainder of this paper is structured as follows. Section 2.1 summarizes the necessary notations and problem settings. Section 2.2 discusses additive noise models and their identifiability conditions. Section 3 introduces a new class of identifiable ANMs, and explains how the models are identifiable. In Sections 3.1, the new identifiability condition in linear settings is restated and compared to the previous identifiability conditions. Section 4 introduces the graph structure learning algorithm, referred to as uncertainty scoring (US). Furthermore, Section 4.1 provides theoretical guarantees for learning Gaussian linear SEMs via the US algorithm. Section 5 explains the numerical experiments and provides an evaluation of the US algorithm against other state-of-the-art DAG learning algorithms, such as the GDS, LISTEN, and LINGAM algorithms when recovering the graphs. Finally, Section 6 compares our algorithm to the GES, GDS, LISTEN, and LINGAM algorithms by analyzing real mathematics marks data.

## 2. Preliminaries

We first introduce some necessary notations and definitions for directed acyclic graphical (DAG) models, additive noise models (ANMs), and linear structural equation models (SEMs). Then, we provide some detailed descriptions of the previous works on the identifiability of ANMs and (Gaussian) linear SEMs in Shimizu et al. (2006); Hoyer et al. (2009); Peters et al. (2012); Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Chen et al. (2019).

### 2.1. Problem Set-up and Notations

A directed acyclic graph $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \cdots, p\}$ and a set of directed edges $E \subset V \times V$ with no directed cycles. A directed edge from node $j$ to $k$ is denoted by $(j, k)$ or $j \to k$. The set of *parents* of node $k$ denoted by $\mathrm{Pa}(k)$ consists of all nodes $j$ such that $(j, k) \in E$. If there is a directed path $j \to \cdots \to k$, then $k$ is called a *descendant* of $j$ and $j$ is an *ancestor* of $k$. The set $\mathrm{De}(k)$ denotes the set of all descendants of node $k$. The *non-descendants* of node $k$ are $\mathrm{Nd}(k) := V \setminus (\{k\} \cup \mathrm{De}(k))$. The *length* of a directed path $j \to j+1 \to ... \to k$ is the number of sequential edges. An important property of DAGs is that there exists a (possibly non-unique) *ordering* $\pi = (\pi_1, ...., \pi_p)$ of a directed graph that represents directions of edges such that for every directed edge $(j, k) \in E$, $j$ comes before $k$ in the ordering. Hence, learning a graph can be decomposed into learning the ordering and the skeleton that is the set of directed edges without their directions.

We consider a set of random variables $X := (X_j)_{j \in V}$ with a probability distribution taking values in probability space $\mathcal{X}_V$ over the nodes in the graph $G$. Suppose that a random vector $X$ has a joint probability density function $P(G) := P(X_1, X_2, ..., X_p)$. For any subset $S$ of $V$, let $X_S := \{X_j : j \in S \subset V\}$ and let $\mathcal{X}(S) := \times_{j \in S} \mathcal{X}_j$. For any node $j \in V$, $P(X_j \mid X_S)$ denotes the conditional distribution of a variable $X_j$ given a random vector $X_S$. With these notations, a joint distribution of a DAG model has the following factorization:

$$P(G) = P(X_1, X_2, ..., X_p) = \prod_{j=1}^{p} P(X_j \mid X_{\mathrm{Pa}(j)}), \tag{1}$$

where $P(X_j \mid X_{\mathrm{Pa}(j)})$ is the conditional distribution of a variable $X_j$ given its parent variables $X_{\mathrm{Pa}(j)} := \{X_k : k \in \mathrm{Pa}(j) \subset V\}$.

Throughout the paper, we assume causal sufficiency such that all relevant variables have been observed. Causal sufficiency is also assumed in the Gaussian SEM learning algorithms of Peters and Bühlmann (2014); Loh and Bühlmann (2014); Ghoshal and Honorio (2017, 2018). However, we assume neither the (adjacent) faithfulness assumption nor bounded number of neighbors or parents assumption that could be very restrictive (Uhler et al., 2013; Park and Park, 2019b). Lastly, for ease of presentation, we sometimes use the node $j \in V$ for the variables $X_j$ as a slight abuse of notation.

### 2.2. Additive Noise Models and their Identifiability

Additive noise models and linear structural equation models, also known as functional models (Pearl, 2014), are a special case of DAG models where the joint distribution is defined by the following structural equations with additive noise: For all $j \in V$,

$$X_j = f_j(X_{\mathrm{Pa}(j)}) + \epsilon_j, \tag{2}$$

where $(f_j)_{j\in V}$ are allowed to have any form of Borel measurable functions and $(\epsilon_j)_{j\in V}$ are independent, but can be different distributions with mean zero and heterogeneous variances $(\sigma_j^2)_{j\in V}$. Here, they are denoted as $\epsilon_j \sim (0, \sigma_j^2)$. Hence, we have $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_p)^T \sim (\mathbf{0}_p, \Sigma_\epsilon)$ where $\mathbf{0}_p = (0, 0, ..., 0)^T \in \mathbb{R}^p$, and $\Sigma_\epsilon$ is a diagonal matrix with unknown variances, which are $\sigma_1^2, \sigma_2^2, ..., \sigma_p^2$.

A linear SEM is a special ANM where $(f_j)_{j\in V}$ are all linear functions, and hence, the joint distribution of a linear SEM is defined by the following linear equations: For all $j \in V$,

$$X_j = \beta_{j0} + \sum_{j' \in \mathrm{Pa}(j)} \beta_{j'j} X_{j'} + \epsilon_j. \tag{3}$$

The linear SEM in Equation (3) can be restated in the following matrix form:

$$(X_1, X_2, ..., X_p)^T = B_0^T + B^T (X_1, X_2, ..., X_p)^T + (\epsilon_1, \epsilon_2, ..., \epsilon_p)^T, \tag{4}$$

where $B_0 \in \mathbb{R}^p$ is an intercept vector, and $B \in \mathbb{R}^{p\times p}$ is an edge weight matrix, an auto regression matrix, or a weighted adjacency matrix with each element $[B]_{jk} = \beta_{jk}$, in which $\beta_{jk}$ is the linear weight of an edge from $X_j$ to $X_k$.

The edge weight matrix $B$ encodes the graph structure under the *non-zero edge weights condition* where $\beta_{j'j}$ is non-zero if $j'$ is a parent of $j$, otherwise, $\beta_{j'j} = 0$, as in other linear structural equation models (see details in Spirtes 1995; Peters and Bühlmann 2014). It is a natural condition in accordance with the intuitive understanding of causal or functional relationships among variables. In linear SEMs, Theorems 1, 2, and 3 in Spirtes (1995) and Lemma 4 in Peters and Bühlmann (2014) show that the non-zero edge weights condition is equivalent to the Markov and *causal minimality* assumptions. Causal minimality means that a joint distribution is not Markov with respect to a strict sub-graph of the true graph. In the settings used here, it is equivalent to the following statement: For any node $j \in V$ and one of its parents $k \in \mathrm{Pa}(j)$,

$$\forall \;\; \mathrm{Pa}(j) \setminus \{k\} \subset S \subset \mathrm{Nd}(j) \setminus \{k\}, \qquad X_j \not\perp\!\!\!\perp X_k \mid X_S.$$

Causal minimality may not be naturally satisfied in a general ANM. Hence, this paper assumes causal minimality, but does not assume faithfulness that is a very strong form of causal minimality. As discussed, faithfulness is commonly assumed for learning a DAG model, such as in the PC (Spirtes et al., 2000) and the max-min hill-climbing (Tsamardinos et al., 2006) algorithms. Nevertheless, in practice, faithfulness cannot be tested, and can be very restrictive in finite sample settings (Uhler et al., 2013).

Without loss of generality, this paper assumes that $\mathbb{E}(X_j) = 0$ for all $j \in V$. Then, the distribution of the linear SEM in Equation (4) is as follows:

$$X \sim (\mathbf{0}_p, \Sigma_X) = \big(\mathbf{0}_p, (I_p - B)^{-1}\Sigma_\epsilon (I_p - B)^{-T}\big),$$

where $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix, and $\Sigma_\epsilon$ is a covariance matrix of the additive errors.

For a special-case Gaussian linear SEM where all error distributions are Gaussian, the density function can be parameterized by the inverse covariance or concentration matrix $\Theta = (I_p - B)^T \Sigma_\epsilon^{-1} (I_p - B) \succ 0$, and can be written as:

$$f_G(x_1, x_2, ..., x_p; \Theta) = \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\Big(-\frac{1}{2}(x_1, ..., x_p)\Theta(x_1, ..., x_p)^T\Big). \tag{5}$$

Due to this convenient form, recent Gaussian linear SEM learning approaches exploit various inverse covariance matrix estimation methods. Loh and Bühlmann (2014) applies graphical Lasso, and Ghoshal and Honorio (2017, 2018) apply the constrained $\ell_1$-minimization for inverse covariance matrix estimation (CLIME).

Regarding to the identifiability, recent works prove the identifiable classes of ANMs by restricting the form of dependency functions $(f_j)_{j \in V}$ (Hoyer et al., 2009) or non-Gaussian error distributions (Hoyer et al., 2009; Mooij et al., 2009; Peters et al., 2012). In addition, if error distributions are Gaussian and dependency functions $(f_j)_{j \in V}$ are all linear, Peters and Bühlmann (2014); Loh and Bühlmann (2014); Ghoshal and Honorio (2017) prove that Gaussian linear SEMs with equal or known error variances are identifiable. More recently, Ghoshal and Honorio (2018); Chen et al. (2019) independently show that (Gaussian) linear SEMs with unknown heterogeneous error variances can be identifiable. We refer the readers to Peters et al. (2014); Eberhardt (2017); Glymour et al. (2019) for details. The following lemma summarizes the identifiable class of ANMs that have been proven.

**Lemma 1 (Identifiable Class of ANMs)** *The following sets have been shown to be identifiable ANMs (2) where $X_j = f_j(X_{Pa(j)}) + \epsilon_j$ for all $j \in V$ :*

- *non-linear ANMs where all $(f_j)_{j \in V}$ are not linear,*
- *non-Gaussian linear ANMs where all $(f_j)_{j \in V}$ are linear, and the distributions of either $(X_j)_{j \in V}$ or $(\epsilon_j)_{j \in V}$ belongs to a set of some non-Gaussian distributions, and*

- *(Gaussian) linear ANMs where all $(f_j)_{j \in V}$ are linear, and the variances of $\epsilon_j$ are the similar or known.*

Detailed proof is provided in Shimizu et al. (2006); Peters et al. (2010); Hoyer et al. (2009); Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Chen et al. (2019). Lemma 1 claims that the underlying graph is recoverable from only the joint distribution if any identifiable assumption is satisfied. In many areas, these classes of ANMs are acceptable and widely used, for example, the assumption of the exact same error variances, proposed in Peters and Bühlmann (2014), is used for applications with variables from a similar domain, spatial or time-series data. However, it might also be unrealistic for real-world data to have exactly the same error variances. In addition, the assumptions about all non-Gaussian error distributions and all non-linear dependency functions might be unrealistic in the same manner.

Therefore, the main focus of this paper is to propose a different identifiability condition for ANMs without constraints on the form of dependency functions and error distributions by applying not only the scale of error variances, but also that of the influence of parents. Not surprisingly, our condition is strictly milder than the equal error variance, but a generalized condition of the recently introduced identifiability conditions for linear SEMs in Ghoshal and Honorio (2018); Chen et al. (2019). We provide the details of our new identifiability condition in the next section.

## 3. Identifiability

This section explains the new identifiability condition for ANMs with any forms of dependency functions and heterogeneous error distributions. To provide intuition, we explain how bivariate Gaussian linear SEM (5) with unknown homogeneous error variances can be identifiable from only the joint distribution illustrated in Figure 1: $G_1 : X_1 = \epsilon_1$ and $X_2 = \beta_1 X_1 + \epsilon_2$, $G_2 : X_1 = \beta_2 X_2 + \epsilon_1$ and $X_2 = \epsilon_2$, and $G_3 : X_1 = \epsilon_1$ and $X_2 = \epsilon_2$, where $\epsilon_j \sim N(0, \sigma_j^2)$ for all $j \in \{1, 2\}$.

Now, we explain how to determine if the underlying graph is either $G_1$, $G_2$, or $G_3$. For $G_1$, if the error variance ratio satisfies $\sigma_2^2/\sigma_1^2 > (1 - \beta_1^2)$, we can see the following two conditions from the law of total variance:

$$\text{(A) } \mathrm{Var}(X_2) = \mathbb{E}(\mathrm{Var}(X_2 \mid X_1)) + \mathrm{Var}(\mathbb{E}(X_2 \mid X_1)) = \sigma_2^2 + \beta_1^2\sigma_1^2 > \sigma_1^2 = \mathrm{Var}(X_1), \quad \text{and}$$

$$\text{(B) } \mathbb{E}(\mathrm{Var}(X_1 \mid X_2)) = \mathrm{Var}(X_1) - \mathrm{Var}(\mathbb{E}(X_1 \mid X_2)) = \sigma_1^2 - \frac{\beta_1^2\sigma_1^4}{\beta_1^2\sigma_1^2 + \sigma_2^2} < \sigma_2^2 = \mathbb{E}(\mathrm{Var}(X_2 \mid X_1)).$$

The former relationship (A) can be understood to mean that the uncertainty level of $X_1$ is lower than the uncertainty level of $X_2$. It intuitively makes sense, because $X_1$ has only one random source $\epsilon_1$, whereas $X_2$ has two random sources, $X_1$ and $\epsilon_2$. Similarly, relationship (B) can be understood to mean that after eliminating the other variable effect, the uncertainty level of $X_1$ is smaller than that of $X_2$. This also makes sense, because the remaining part of $X_2$ after eliminating the effect of $X_1$ is $\epsilon_2$ , while the remaining part of $X_1$ is a part of $\epsilon_1$ since $X_2$ contains some information of $\epsilon_1$. Hence, even when the error variances are different, we can recover the ordering as long as $\sigma_2^2/\sigma_1^2 > (1 - \beta_1^2)$.

In the same manner, we can find the true ordering $\pi = (2, 1)$ for $G_2$ as long as $\sigma_1^2/\sigma_2^2 > (1 - \beta_2^2)$. Lastly, for $G_3$, there is no guarantee as to which marginal or conditional variance

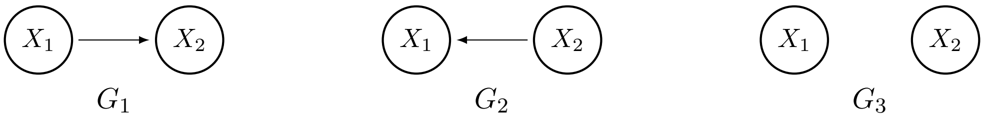


Figure 1: Bivariate directed acyclic graphs of $G_1$, $G_2$, and $G_3$

is bigger since $(1, 2)$ and $(2, 1)$ are correct orderings of $G_3$. Hence, any choice is correct. Therefore, we can recover the orderings of $G_1$, $G_2$, and $G_3$ by testing which marginal or conditional variance is bigger.

The presence of an edge is easily verified from the dependence relationships between variables. For $G_1$ and $G_2$, $X_1$ and $X_2$ are dependent, whereas for $G_3$, $X_1$ and $X_2$ are independent. Therefore, combined with the ordering, we can recover the true graph.

Now, we extend this to general $p$-variate ANMs with any forms of dependency functions and heterogeneous error distributions with unknown variances. Since both the law of total variance and independence relationships do not require linearity of $f$ and Gaussian error distributions, the key idea to extending model identifiability from the bivariate to the multivariate still involves the comparisons of the (conditional) node variances.

**Theorem 2 (Identifiability Conditions for ANMs)** *Let $P(X)$ be generated from an ANM (2) with DAG $G$ and true ordering $\pi$. Suppose that causal minimality holds. Then, DAG $G$ is uniquely identifiable if either of the two following conditions is satisfied: For any node $j = \pi_m \in V$, $k \in De(j)$, and $\ell \in An(j)$,*

*(A) Forward stepwise selection: $\sigma_j^2 < \sigma_k^2 + \mathbb{E}(Var(\mathbb{E}(X_k \mid X_{Pa(k)}) \mid X_{\pi_1}, ..., X_{\pi_{m-1}}))$, or*

*(B) Backward stepwise selection: $\sigma_j^2 > \sigma_\ell^2 - \mathbb{E}(Var(\mathbb{E}(X_\ell \mid X_{\pi_1}, ..., X_{\pi_m} \setminus X_\ell) \mid X_{Pa(\ell)}))$.*

The detailed proof is provided in the Appendix A. Theorem 2 claims that ANMs are identifiable if either the conditional variance of a node $j$ is smaller than that of its descendant, De$(j)$, given the non-descendant, Nd$(j)$, or if the conditional variance of a node $j$ given its parents is bigger than that of its ancestor, An$(j)$, given the union of its parents and any of its descendants. The former condition can be understood to mean that the variance of $\epsilon_j$ is overestimated owing to lack of parents. And, the latter can be understood to mean that the estimated variance of $\epsilon_j$ is smaller than the true variance due to all parents plus the addition of descendants that can explain $\epsilon_j$. As shown later in Algorithm 3, Condition (A) is applied to the element-wise selection of the ordering starting from the first one, $\pi_1$, and Condition (B) is used for the component-wise selection of the ordering starting from the last one, $\pi_p$.

We note that both Conditions (A) and (B) are immediately satisfied when the error variances are the same, which is the identifiability assumption for linear SEMs in Peters and Bühlmann (2014); Loh and Bühlmann (2014); Ghoshal and Honorio (2017); Chen et al. (2019). More generally, we obtain the following sufficient condition for both Conditions (A) and (B).

**Corollary 3** *Consider an ANM (2) with DAG $G$. If the error variances are the same or weakly monotone increasing in the ordering, then DAG $G$ is uniquely identifiable.*

An important remaining question is the relationship between Conditions (A) and (B) in Theorem 2. Note that Conditions (A) and (B) are equivalent for any bivariate graphs as presented in Figure 1. However, they are, in general, not always equivalent. We investigate the relationship between Conditions (A) and (B) using a 3-node linear SEM in the next section.

### 3.1. Identifiability of Linear Structural Equation Models

In the linear SEM (3) setting, a conditional variance can be obtained from the inverse covariance matrix, $\Theta = (I_p - B)^T \Sigma_\epsilon^{-1}(I_p - B)$. Hence, our new identifiability conditions in Theorem 2 can be expressed with error variances as well as edge weights, and we have the following identifiability conditions.

**Theorem 4 (Identifiability Conditions for Linear SEMs)** *Let $P(X)$ be generated from a linear SEM (3) with DAG $G$ and true ordering $\pi$. Then, DAG $G$ is uniquely identifiable, if either of the two following conditions is satisfied: for any node $j \in V$, $k \in De(j)$ and $\ell \in An(j)$,*

*(A) Forward stepwise selection:*

$$\sigma_j^2 < \sigma_k^2 + \sum_{k' \in Pa(k) \setminus \{\pi_1, \ldots, \pi_{m-1}\}} \beta_{k' \to k}^2 \sigma_{k'}^2$$

*where $j = \pi_m$ and $\beta_{k' \to k}$ is the sum over products of coefficients along each directed paths from $k'$ to $k$ of products of coefficients along each path.*

*(B) Backward stepwise selection:*

$$\frac{1}{\sigma_j^2} < \frac{1}{\sigma_\ell^2} + \sum_{\ell' \in Ch(\ell) \setminus \{\pi_m, \ldots, \pi_p\}} \frac{\beta_{\ell\ell'}^2}{\sigma_{\ell'}^2}.$$

The detailed proof is provided in Appendix B. Theorem 4 claims that the underlying graph can be uniquely recoverable even when error variances are different. Specifically, Condition (A) exploits the sum of squares of all the influences from non-considered parents in the condition set, whereas Condition (B) applies the sum of squares only of direct influences from a node to its child.

Condition (A) can be derived from Chen et al. (2019) although it only focuses on a Gaussian linear SEM with the equal variances. Furthermore, Condition (B) is the same as the identifiability condition for linear SEMs in Ghoshal and Honorio (2018). Hence, our identifiability result can be understood as a generalized version of recent linear SEM identifiability conditions. However, we emphasize that our new conditions have never been proposed for general ANMs with unknown heterogeneous error variances.

Lastly, we investigate the relationship between Conditions (A) and (B) in Theorem 4 using a simple 3-node chain graph. Consider a linear SEM, $X_1 \to X_2 \to X_3$ such that $X_1 = \epsilon_1$, $X_2 = \beta_1 X_1 + \epsilon_2$, and $X_3 = \beta_2 X_2 + \epsilon_3$ where $\epsilon_j \sim N(0, \sigma_j^2)$ for all $j \in \{1, 2, 3\}$. Then, Condition (A) in Theorem 4 is equivalent to the following three conditions:

$$(A1)\ \ \sigma_1^2 < \sigma_2^2 + \beta_1^2\sigma_1^2, \quad (A2)\ \ \sigma_2^2 < \sigma_3^2 + \beta_2^2\sigma_2^2, \quad (A3)\ \ \sigma_1^2 < \sigma_3^2 + \beta_2^2\sigma_2^2 + \beta_1^2\beta_2^2\sigma_1^2.$$

**Algorithm 1: Ordering estimation using the forward stepwise selection**

```
Input   : $n$ i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated ordering, $\widehat{\pi} = (\widehat{\pi}_1, ..., \widehat{\pi}_p)$

Set $\widehat{\pi}_0 = \emptyset$
for $m = \{1, 2, \cdots, p\}$ do
    Set $S = \{\widehat{\pi}_0, ..., \widehat{\pi}_{m-1}\}$
    for $j \in \{1, 2, \cdots, p\} \setminus S$ do
        Estimate the conditional variance of $X_j$ given $X_S$, $\widehat{\sigma}^2_{j|S}$
    end
    The $m$-th element of the ordering $\widehat{\pi}_m = \arg\min_j \widehat{\sigma}^2_{j|S}$
end
```

In contrast, Condition (B) is equivalent to the following three conditions:

$$(B1)\ \frac{\sigma_2^2}{\sigma_1^2} > (1-\beta_1^2), \quad (B2)\ \frac{\sigma_3^2}{\sigma_2^2} > (1-\beta_2^2), \quad (B3)\ \frac{\sigma_3^2}{\sigma_1^2} > 1 - \frac{\beta_1^2\sigma_3^2}{\sigma_2^2}.$$

As shown in Corollary 3, when error variances are monotone increasing (that is, $\sigma_1^2 \leq \sigma_2^2 \leq \sigma_3^2$), both conditions are always satisfied. Hence, we consider a case where the error variances are strictly monotone decreasing, that is, $\sigma_j^2 = a\sigma_{j-1}^2$ for some $0 < a < 1$. Then, simple algebra yields that Condition (A) is equivalent to $\beta_1^2 > 1 - a$ and $\beta_2^2 > 1 - a$. In addition, Condition (B) is equivalent to $\beta_1^2 > \frac{1-a^2}{a}$ and $\beta_2^2 > 1 - a$. Hence, in this setting, Condition (A) is strictly milder.

We also consider another case where $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (2, 2, 1)$ and $\beta_2 = 1$. In this setting, Condition (B) is violated if $\beta_1^2 \leq 1$, while Condition (A) always holds. However, in a different case where $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (2, 1, 1)$ and $\beta_1 = 1$, Condition (A) is violated if $\beta_2^2 \leq 1/3$, while Condition (B) always holds. Therefore, we again cannot conclude that one is strictly weaker than another, in general.

## 4. Algorithms

In this section, we present the uncertainty scoring (US) algorithm (Algorithm 3) for learning our new class of identifiable ANMs based on the forward and backward stepwise selection conditions in Theorem 2. The US algorithm consists of two steps: (1) element-wise ordering estimation from either the initial or terminal using the conditional variances; and (2) parent estimation using the conditional independence relationships between variables. For each step of the US algorithm, any conditional variance estimation method and any independence test can be applied.

Regarding the ordering estimation in Step (1), Algorithms 1 and 2 require the conditional variance of each variable. Hence, we can use a consistent estimator for the error variances using any regression methods, such as ordinary linear regression, regularized regression, generalized additive model regression, and local polynomial regression as in Nowzohour and Bühlmann (2016). For an example of the conditional variance $\mathrm{Var}(X_j \mid X_S)$ in a linear

**Algorithm 2: Ordering estimation using the backward stepwise selection**

```
Input  : $n$ i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated ordering, $\widehat{\pi} = (\widehat{\pi}_1, ..., \widehat{\pi}_p)$
Set $S = \{1, 2, \cdots, p\}$
for $m = \{p, p-1, \cdots, 1\}$ do
    for $j \in S$ do
        Estimate the conditional variance $X_j$ given $X_{S \setminus j}$, $\widehat{\sigma}^2_{j|S \setminus j}$
    end
    The $m$-th element of the ordering $\widehat{\pi}_m = \arg\max_j \widehat{\sigma}^2_{j|S \setminus j}$
    Update $S = S \setminus \pi_m$
end
```

**Algorithm 3: Uncertainty Scoring (US) algorithm**

```
Input  : $n$ i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated directed acyclic graph, $\widehat{G} = (V, \widehat{E})$
Step (1): Ordering Estimation
Estimate the ordering $\widehat{\pi}$ using Algorithm 1 or 2
Step (2): Parents Estimation
for $m = \{2, \cdots, p\}$ do
    for $j = \{1, \cdots, m-1\}$ do
        Perform a conditional independence test between $\widehat{\pi}_m$ and $\widehat{\pi}_j$ given
          $\{\widehat{\pi}_1, ..., \widehat{\pi}_{m-1}\} \setminus \widehat{\pi}_j$
        If dependent, include $\widehat{\pi}_j$ into $\widehat{\text{Pa}}(\widehat{\pi}_m)$
    end
end
Estimate the edge set $\widehat{E} := \cup_{m \in \{2,3,...,p\}} \cup_{k \in \widehat{\text{Pa}}(\widehat{\pi}_m)} (k, \widehat{\pi}_m)$
```

SEM, first regress $X_j$ over $X_S$, and then, estimate $\text{Var}(X_j \mid X_S)$ using its residuals. We describe this more precisely with its statistical guarantees in Section 4.1.

Under Condition (A) in Theorem 2, the conditional variance of the correct element of the ordering $\pi_j$ given $\pi_1, ..., \pi_{j-1}$ is strictly smaller than that of the other nodes in the population. Hence, Algorithm 1 can find the correct element of the ordering that has the smallest conditional variance. For the next element of the ordering $\pi_{j+1}$, we compute all conditional variances given $\pi_1, ..., \pi_j$ and choose the node with the smallest conditional variance. Therefore, Algorithm 1 learns the ordering from the beginning by selecting the node with the minimum conditional variance and updating the condition set.

In the same manner, under Condition (B) in Theorem 2, the conditional variance of the correct element of the ordering $\pi_j$ given $V \setminus \{\pi_j, \pi_{j+1}, ..., \pi_p\}$ is strictly bigger than that of any node $\pi_k$ given $V \setminus \{\pi_k, \pi_{j+1}, ..., \pi_p\}$ for $k \in \{1, 2, ..., j-1\}$ in the population. Hence, we can also determine one node at a time by selecting the node with the largest conditional variance, and hence, we can recover the true ordering from the last using Algorithm 2.

Estimating the parents of a node $\pi_j$ in Step (2) of Algorithm 3 is equivalent to selecting the parents among all the elements before a node $\pi_j$ in the ordering. Hence, given the estimated ordering from Step (1), Step (2) is reduced to a neighborhood selection problem using conditional dependence relationships like the constraint-based graph structure learning PC algorithm. More precisely, the parent of a node $\pi_j$ is determined as a set of nodes $k$ such that $\pi_j$ and $k$ are conditionally dependent given $\{\pi_1, ..., \pi_{j-1}\} \setminus \{k\}$, that is $\text{Pa}(\pi_j) := \{k \in \{\pi_1, ..., \pi_{j-1}\} \mid X_{\pi_j} \not\perp X_k \mid \{X_{\pi_1}, ..., X_{\pi_{j-1}}\} \setminus X_k\}$. However, unlike the PC algorithm, our approach does not require faithfulness and a greedy search, owing to the estimated ordering in Step (1).

When focusing on learning a linear SEM, the main strategy of Algorithms 3 is analogous to the algorithms using the inverse covariance matrix, as done by Loh and Bühlmann (2014); Ghoshal and Honorio (2017, 2018); Chen et al. (2019), where each element of the ordering is estimated from the end, and then, its parent is estimated. However, the proposed algorithm focuses on learning general ANMs regardless of the complexity degree of a graph, whereas their algorithms are for learning sparse linear SEMs in high dimensional settings. Hence, when the algorithms of Loh and Bühlmann (2014); Ghoshal and Honorio (2017, 2018); Chen et al. (2019) fail to recover the true graph due to a violation of the linearity assumption, Algorithm 3 can recover the true underlying graph. In addition, our algorithm does not require the faithfulness assumption, whereas the algorithms in Loh and Bühlmann (2014); Ghoshal and Honorio (2017, 2018) require the (adjacent) faithfulness assumption.

In terms of the computational complexity, Algorithm 3 involves $O(p^2)$ estimations of conditional variances in Step (1) and $O(p^2)$ conditional independence tests in Step (2). However, the detailed computational complexity of each step relies on the choice of an estimation method. For a special case of learning a Gaussian linear SEM, the inverse of a sample covariance matrix can be applied to both Steps (1) and (2). Then, we can see that the worst-case computational complexity of Algorithm 3 is $O(np^5)$.

We empirically verify that Algorithms 1 and 2 successfully recover sparse and non-sparse linear SEMs with homogeneous as well as heterogeneous error variances in Sections 5.1 and 5.2, respectively. Also shown through simulations in Sections 5.3 and 5.4 is that non-linear ANMs and non-Gaussian ANMs are successfully recovered by the proposed algorithms.

### 4.1. Theoretical Guarantees for Learning Gaussian Linear SEMs

This section provides theoretical guarantees on each step of our algorithms for learning a Gaussian linear SEM when the ordinary least squares method is applied in Step (1), and Fisher's z-transform of the partial correlation is exploited in Step (2). The main result is expressed in terms of the sample size and the fixed node size of the graph.

We begin by discussing the assumptions we impose on Gaussian linear SEMs. Since the ordinary least squares approach and partial correlations are applied, the assumptions involve the covariance matrix and partial correlations. The first assumption is that the minimum and maximum eigenvalues of the covariance matrix are bounded.

**Assumption 5** *Let $X$ be generated from a linear SEM (3). There exist positive constants $\rho_{min}$ and $\rho_{max}$ such that the smallest and largest eigenvalue of covariance matrix, $\Sigma = \text{Cov}(X)$, are bounded.*

$$\rho_{min} \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq \rho_{\max},$$

*where $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ are the smallest and largest eigenvalues of matrix $A$, respectively.*

This assumption can be understood as ensuring that variables are not overly dependent. Also required is the following assumption on partial correlations, to ensure that conditional independent tests using partial correlations are consistent, as discussed in Kalisch and Bühlmann (2007).

**Assumption 6 (Bounded Partial Correlations)** *Let $X$ be generated from a linear SEM* (3)*. For any edge $(k, j) \in E$ and $Pa(j) \setminus \{k\} \subset S \subset Nd(j) \setminus \{k\}$, there exists $M > 0$ and $\kappa(n) = O(\sqrt{1/\log(n)})$ such that*

$$0 < \kappa(n) < \inf_{j,k,S} |\rho_{j,k,S}| < \sup_{j,k,S} |\rho_{j,k,S}| < M < 1,$$

*where $\rho_{j,k,S}$ is the partial correlation between $X_j$ and $X_k$ given $X_S$.*

Lastly, a stronger version of the identifiability assumption in Theorem 4 is required because we move from the population to finite samples.

**Assumption 7** *Consider a linear SEM* (3) *with a true ordering $\pi = (\pi_1, ..., \pi_p)$. For any node $j = \pi_m \in V$, $k \in De(j)$, and $\ell \in An(j)$, there exist positive constants $\tau_F$ and $\tau_B$ such that*

- *Forward stepwise selection:*

$$\sigma_j^2 + \tau_F < \sigma_k^2 + \sum_{k' \in Pa(k) \setminus \{\pi_1, ..., \pi_{m-1}\}} \beta_{k' \to k}^2 \sigma_{k'}^2, \quad \text{or}$$

- *Backward stepwise selection:*

$$\frac{1}{\sigma_j^2} - \tau_B < \frac{1}{\sigma_\ell^2} + \sum_{\ell' \in Ch(\ell) \setminus \{\pi_m, ..., \pi_p\}} \frac{\beta_{\ell\ell'}^2}{\sigma_{\ell'}^2}.$$

Applying Assumptions 5, and 7, we have the following main result whereby the ordering can be successfully recovered via Algorithms 1 and 2.

**Theorem 8 (Recovery of the Ordering)** *Consider a linear SEM* (3) *with a true set of orderings $\Pi$. Suppose that $n > p$ where $p$ is the number of nodes, and $\widehat{\pi}$ is the estimated ordering via Algorithm 1 or 2. Then, there exist positive constants $C_1$ and $C_2$ such that*

$$\mathbb{P}(\widehat{\pi} \in \Pi) \geq 1 - C_1 p^2 \; exp\left(-C_2 \frac{n}{\log n}\right).$$

The detailed proof is in Appendix C. Theorem 8 shows that both Algorithms 1 and 2 consistently recover the ordering of a linear SEM. The condition $n > p$ is necessary because ordinary linear regression is applied. However, it can be relaxed if other conditional variance estimation methods are applied for high dimensional settings. This is also left for future work.

Under Assumption 6, another main result is reached, such that the true directed edges of Gaussian linear SEM can be recovered via our algorithm.

**Theorem 9 (Recovery of the Directed Edges)** *Consider a Gaussian linear SEM* (5). *Suppose that true ordering is provided and let $\widehat{E}$ be the estimated edges from Algorithm 3 with significance level $\alpha(n) = 2(1-\Phi(\sqrt{n}\cdot\kappa(n)/2))$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Then, there exist positive constants $C_3$ and $C_4$ such that*

$$\mathbb{P}(\widehat{E} = E) \geq 1 - C_3 np^2 \; exp\left(-C_4 n \cdot \kappa(n)^2\right).$$

We provide the detailed proof in Appendix D. Theorem 9 claims that when the sample size is sufficiently large, comparing to the lower bound of the true partial correlations, Step 2 of the US algorithm can recover the parent of each node with a high probability.

Finally, by combining Theorems 8 and 9, we reach our final main result that our algorithm successfully recovers the true structure of a Gaussian linear SEM with a high probability.

**Corollary 10 (Recovery of the Graph Structure)** *Consider a Gaussian linear SEM* (5). *Suppose that $n > p$, and Assumptions 5, 6, and 7 are satisfied. Let $\widehat{G} = (V, \widehat{E})$ be the estimated graph from Algorithm 3. Then, there exist positive constants $D_1$ and $D_2$ such that*

$$\mathbb{P}(\widehat{G} = G) \geq 1 - D_1 np^2 \; exp\left(-D_2 \frac{n}{\log n}\right).$$

Provided so far are sample complexity guarantees of the US algorithm for a Gaussian linear SEM (5) when ordinary linear regression and Fisher's z-transform of the partial correlation are applied. As we discussed, the US algorithm can make use of any appropriate regression method and independence test. Hence, if another learning method is applied for each step, not only Gaussian linear SEMs but general ANMs can be recovered with a high probability. Moreover, one can find the statistical guarantees for a choice of methods for each step.

## 5. Numerical Experiments

This section provides numerical experiments to support our theoretical results: ANMs with non-equal error variances can be identifiable; and Algorithm 3 consistently recovers Gaussian linear SEMs. Hence, considered are (i) Gaussian linear SEMs with equal error variances and (ii) heterogeneous error variances, (iii) Gaussian polynomial SEMs, and (iv) linear SEMs with Gaussian and non-Gaussian error distributions.

The proposed forward and backward selection-based algorithms and the comparison GDS, LISTEN, and LINGAM algorithms are evaluated in terms of the average precision ($\frac{\text{\# of correctly estimated edges}}{\text{\# of estimated edges}}$) and the average recall ($\frac{\text{\# of correctly estimated edges}}{\text{\# of true edges}}$). In addition, we report the Hamming distance between the estimated and true DAGs (# of edges that are different between two graphs). For precision and recall, the bigger the better, but for the Hamming distance, smaller is better. We also provide oracles where the edges for true graph are used while the ordering is estimated via Algorithms 1 and 2. Hence, we can verify how accurately the proposed algorithms recover the true orderings of graphs.

To validate Theorems 8 and 9, ordinary linear regression for Step (1) is applied. In addition, Step (2) of Algorithm 3 were implemented using a Fisher's independence test. We always set the significance level depending on the sample size, $\alpha = 1 - \Phi(n^{1/4}/2)$, as in Theorem 9. For the GDS algorithm, we set the initial graph to a random graph. Since the GDS algorithm uses a greedy search, and its accuracy relies heavily on the initial graph, the GDS algorithm can recover the graph better with an appropriate choice of an initial graph. We do not apply the GDS algorithm to large-scale graphs, $p = 200$, due to the heavy computational cost.

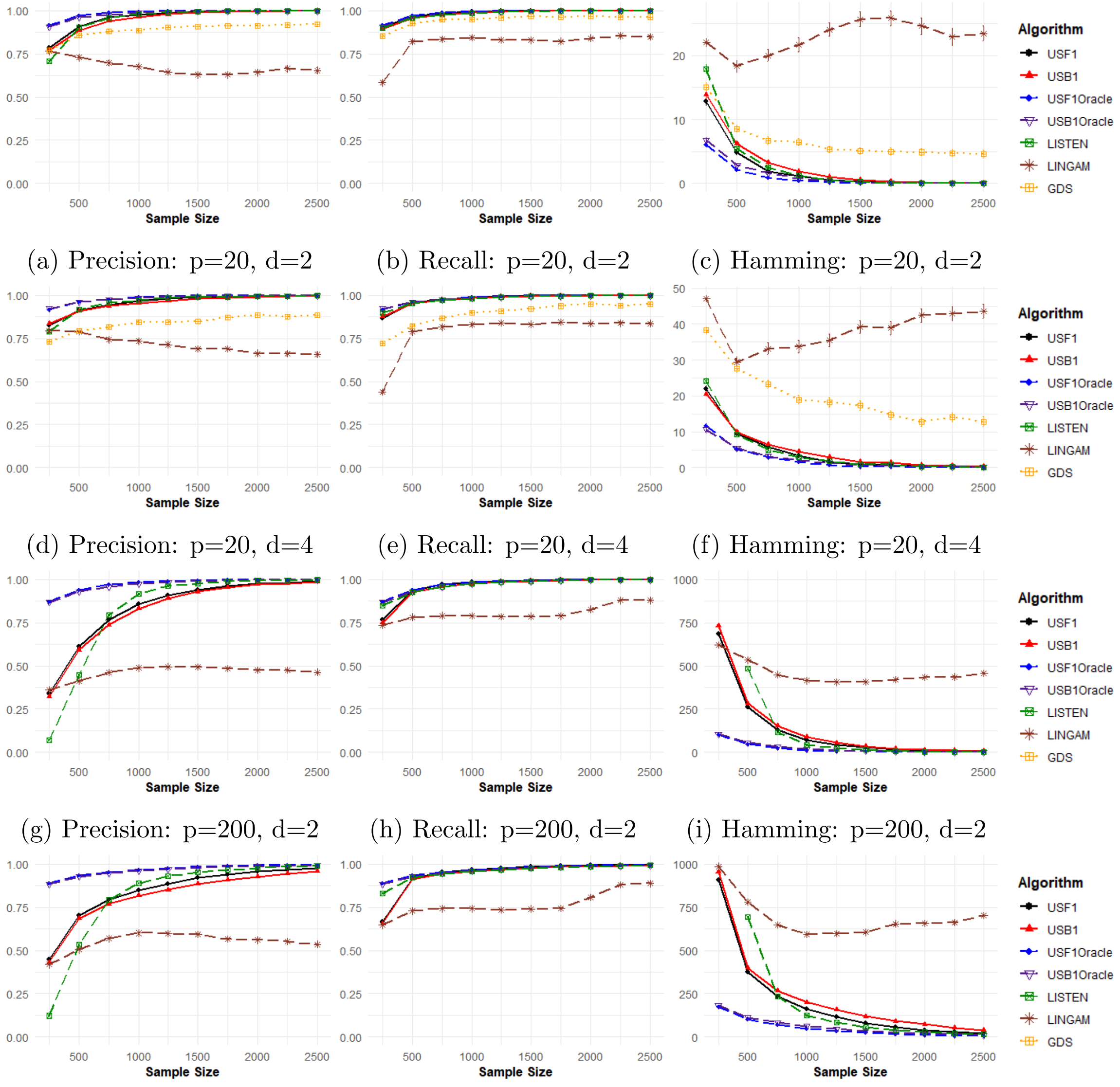
We also emphasize that the LISTEN algorithm of Ghoshal and Honorio (2018) using constrained $\ell_1$-minimization for inverse covariance matrix estimation may perform well with the appropriate choice of regularization parameters. However, when the regularization and hard thresholding parameters are chosen from 10-fold cross validation, the performance in recovering the graph is poor because the cross-validation does not generally have consistency properties for model selection (see details in Shao, 1993). In addition, the algorithm often fails to implement due to the failure of updating the inverse covariance matrix in our settings. Hence, the regularization parameters were set to 0.001, and the hard threshold parameter to half of the minimum value of true edge weights, $\min(|\beta_{jk}|/2)$, by using the true model information because it seems to be much better than the parameters from cross validation when recovering graphs. However, we also point out that our choice of the parameters are not the best, especially when sample size is small. Hence, for a better presentation, we do not present the hamming distance of the LISTEN algorithm for the case of $p = 200, n = 250$ because it is comparatively too large.

As discussed, our algorithms are for inferring the proposed identifiable ANMs using the forward and backward selection conditions. In addition, the GDS algorithm is for learning Gaussian linear SEMs with equal error variances. The LISTEN algorithm is for learning linear SEMs with heterogeneous error variances using the backward selection condition. Lastly, the LINGAM algorithm is designed for learning non-Gaussian linear SEMs. In other words, these algorithms do not guarantee recovering the true graph if the required conditions are not satisfied. Through the numerical experiments, we also point out that the proposed and the comparison algorithms sometimes fail to recover a graph because of the violation of the required assumptions.

### 5.1. Random Gaussian Linear SEMs with Homogeneous Error Variances

We conducted simulations using 100 realizations of $p$-node Gaussian linear SEMs (5) with the randomly generated underlying DAG structures for node size $p \in \{20, 200\}$ while respecting the indegree constraint $d \in \{1, 2\}$. The set of non-zero parameters $\beta_{jk} \in \mathbb{R}$ in Equation (5) were generated uniformly at random in the range $\beta_{jk} \in (-\sqrt{0.5/d}, -\sqrt{0.125}) \cup (\sqrt{0.125}, \sqrt{0.5/d})$. Lastly, all noise variances were set to $\sigma_j^2 = 0.75$. In this setting, we have verified that across hundreds sets of randomly generated samples, all variables have similar marginal variances varying between 0.75 to approximately 2. In addition, we cannot see any patterns between the variances.

Figure 2 evaluates the proposed algorithms and state-of-the-art GDS, LISTEN, and LINGAM algorithms in terms of recovering DAGs by varying sample size $n \in \{250, 500, ..., 2500\}$. Figure 2 also provides oracles where the true skeleton is used while the ordering is estimated via the proposed algorithms, referred to as USF1Oracle and USB1Oracle. Hence,



(a) Precision: p=20, d=2 (b) Recall: p=20, d=2 (c) Hamming: p=20, d=2

(d) Precision: p=20, d=4 (e) Recall: p=20, d=4 (f) Hamming: p=20, d=4

(g) Precision: p=200, d=2 (h) Recall: p=200, d=2 (i) Hamming: p=200, d=2

(j) Precision: p=200, d=4 (k) Recall: p=200, d=4 (l) Hamming: p=200, d=4

Figure 2: Comparison of the proposed algorithms (USF1, USB1), the proposed algorithms with true skeletons (USF1Oracle, USB1Oracle), the GDS, LISTEN, and LINGAM algorithms in terms of average precision, recall, and Hamming distance (Hamming) when recovering Gaussian linear SEMs with homogeneous error variances.

USF1Oracle and USB1Oracle show the accuracy in the recovery of the orderings using the forward and backward selection conditions in Assumption 7, respectively.

As seen in Figure 2, our forward and backward stepwise selection-based algorithms, referred to as USF1 and USB1, recover the true directed edges better as sample size increases and their hamming distances converge to 0. This confirms that Gaussian linear SEMs with homogeneous errors are identifiable, and our algorithms are consistent as proven in Theorems 8 and 9. In addition, we can see that our algorithms perform better for the sparse ($d = 2$) graphs than the dense graphs ($d = 4$). That is mainly because our choice of significance level is more favorable to the sparse graph setting. Another reason is that a sparse graph is more likely to have a larger number of true orderings than a dense graph, and hence, our algorithms are more likely to find the true ordering of a sparse graph.

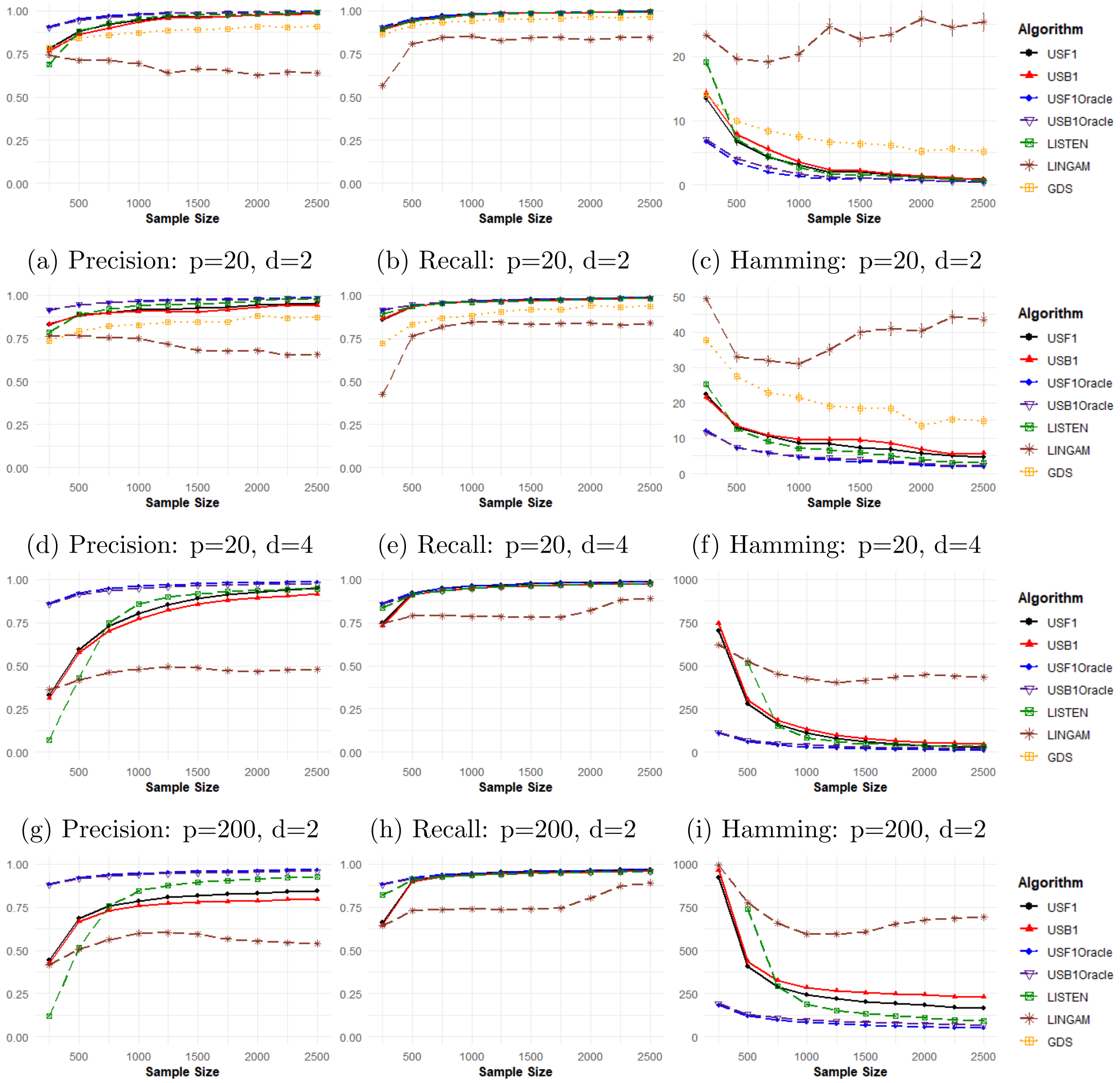
As discussed in Section 3, there is no relationship between the forward and backward stepwise selection conditions. Hence, it is impossible to claim that the forward stepwise selection condition is weaker than the backward stepwise selection condition. Nonetheless, in limited sample settings, Figure 2 shows that USF1Oracle requires slightly fewer samples than USB1Oracle, when the ordering of a graph is estimated. This phenomenon makes sense, because when the first element of the ordering is estimated, the forward selection condition-based Algorithm 1 requires the marginal variances of all nodes. However, the backward selection condition-based Algorithm 2 requires the conditional variances of each node given all other nodes. Since the marginal variance estimation is much easier than the conditional variance estimation in the limited sample settings, the probability that Algorithm 1 chooses the first element of the ordering from among $p$ nodes is higher than the probability that Algorithm 2 selects the last element of the ordering from among $p$ nodes. With the same reasoning, Algorithm 1 may not suffer from the lack of samples when learning the first few elements of the ordering even in high dimensional settings. This phenomenon can also be seen in all other simulation settings in Sections 5.2, 5.3, and 5.4.

Figure 2 shows that the algorithms proposed in this paper generally outperform the comparison GDS algorithm, on average, even with the same error variances, because our method is a complete search-based and exploits the weaker identifiability assumption in Theorem 4. As expected, the LISTEN algorithm performs well when sample size is sufficiently large ($n \geq 500$), and the LINGAM algorithm cannot learn Gaussian linear SEMs. The phenomenon that LISTEN performs better than USF1 and USB1 is not a contradictory result, because LISTEN is designed for learning Gaussian linear SEMs and the hard thresholding parameter is chosen by the true model information. In addition, decreasing accuracy of LINGAM makes sense because the required non-Gaussian assumption is more likely to be violated as sample size increases. However, it does not mean that LINGAM recovers Gaussian linear SEMs better as sample size decreases. As seen in Figure 2, the minimum hamming distances are not achieved when sample size is smallest. Lastly, since the skeleton estimation is not perfect, we can see that USF1 and USB1 are significantly worse than USF1Oracle and USB1Oracle.

### 5.2. Random Gaussian Linear SEMs with Heterogeneous Error Variances

In order to authenticate the validation of the theoretical results that Gaussian linear SEMs with non-equal error variances can be identifiable, we also generated 100 sets of samples under the same procedure specified in Section 5.1, except for randomly chosen error variances, $\sigma_j^2 \in [0.70, 0.80]$. Then, our algorithms and the comparison methods were evaluated by varying sample size $n \in \{250, 500, ..., 2500\}$ for $p \in \{20, 200\}$ in Figure 3.

As expected, most of the simulation results are analogous to the Gaussian linear SEMs with homogeneous error variances in Section 5.1. More precisely, Figure 3 shows that our algorithms, USF1 and USB1, consistently recover the graphs. This heuristically confirms our theoretical findings that ANMs with heterogeneous error variances are identifiable, and



(a) Precision: p=20, d=2 (b) Recall: p=20, d=2 (c) Hamming: p=20, d=2

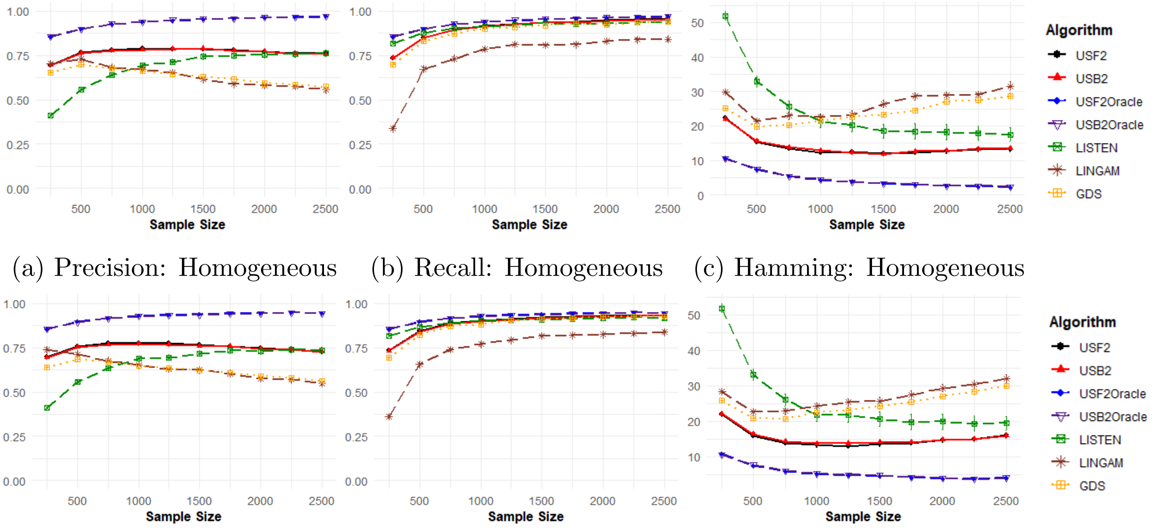
(d) Precision: p=20, d=4 (e) Recall: p=20, d=4 (f) Hamming: p=20, d=4

(g) Precision: p=200, d=2 (h) Recall: p=200, d=2 (i) Hamming: p=200, d=2

(j) Precision: p=200, d=4 (k) Recall: p=200, d=4 (l) Hamming: p=200, d=4

Figure 3: Comparison of the proposed algorithms (USF1, USB1), the proposed algorithms with true skeletons (USF1Oracle, USB1Oracle), the GDS, LISTEN, and LINGAM algorithms in terms of average precision, recall, and Hamming distance (Hamming) for recovering Gaussian linear SEMs with different error variances.

that our algorithms are consistent. In addition, the phenomenon that USF1Oracle performs better than USB1Oracle is more exaggerated when the number of node is large, $p = 200$. Hence, Figure 3 reveals the advantages of the forward stepwise selection approach when recovering large-scale DAG models with heterogeneous error variances. However, for the case of $d = 4$, neither USF1Oracle and USB1Oracle have 0 hamming distance even when sample size is huge. This is mainly because the identifiability conditions in Theorem 3 are often violated in our setting.



(a) Precision: Homogeneous (b) Recall: Homogeneous (c) Hamming: Homogeneous

(d) Precision: Heterogeneous (e) Recall: Heterogeneous (f) Hamming: Heterogeneous

Figure 4: Comparison of the proposed algorithms (USF2, USB2), the proposed algorithms with true skeletons (USF2Oracle, USB2Oracle), the GDS, LISTEN, and LINGAM algorithms in terms of average precision, recall, and Hamming distance (Hamming) for recovering 20-node Gaussian polynomial SEMs with homogeneous and heterogeneous error variances.

Lastly, Figure 3 shows that the GDS algorithm recovers graphs better as a sample size increases. That is also expected, since Peters and Bühlmann (2014) empirically shows that the GDS algorithm possibly learns Gaussian linear SEMs with different error variances without theoretical reasons. We can also see this robustness of equal error variances assumption in Loh and Bühlmann (2014); Ghoshal and Honorio (2017), where they assume the same error variances.

### 5.3. Random Polynomial SEMs with Homogeneous and Heterogeneous Error Variances

A popular non-linear ANM is a polynomial SEM in which each variable is modeled as an m-th degree polynomial in its parents, and hence, it is a non-linear when $m > 1$. More precisely, m-th degree polynomial SEMs have the following form. For all $j \in V$,

$$X_j = \sum_{k \in \mathrm{Pa}(j)} \beta_{1kj} X_k + \sum_{k \in \mathrm{Pa}(j)} \beta_{2kj} X_k^2 + \cdots + \sum_{k \in \mathrm{Pa}(j)} \beta_{mkj} X_k^m + \epsilon_j. \tag{6}$$

For a special case where each error distribution is Gaussian, $\epsilon_j \sim N(0, \sigma_j^2)$, the joint density is as follows.

$$f_G(x_1, x_2, ..., x_p) = \exp\Big( - \sum_{j \in V} \log(\sqrt{2\pi}\sigma_j) - \sum_{j \in V} \frac{1}{2\sigma_j^2} \big(x_j - \sum_{k \in \mathrm{Pa}(j)} \sum_{h \in \{1,...,m\}} \beta_{hkj} X_k^h \big)^2 \Big).$$

To validate the theoretical results that non-linear additive noise models can be identifiable, we conducted simulations using 100 realizations of 20-node Gaussian polynomial

SEMs (6). As in Section 5.1, the randomly generated underlying DAG structures while respecting the indegree constraint $d = 2$ was considered, and we set the maximum degree $m = 2$. The set of non-zero parameters $\beta_{1jk}, \beta_{2jk} \in \mathbb{R}$ in Equation (6) were generated uniformly at random in the range $\beta_{1jk} \in (-0.4, -0.2) \cup (0.2, 0.4)$ and $\beta_{2jk} \in \{-0.2, 0, 0.2\}$. For homogeneous error variances, all noise variances were set to $\sigma_j^2 = 0.5$, and for heterogeneous error variances, all noise variances were randomly chosen $\sigma_j^2 \in [0.475, 0.525]$. We again verified that the generated samples have similar variances in this setting.

Our algorithms were implemented using a quadratic model instead of a linear model for the estimation of error variances. The quadratic models-based algorithms using forward and backward stepwise selections are referred to as USF2 and USB2, respectively. We evaluate the USF2, USB2, GDS, LISTEN, and LINGAM algorithms varying sample size $n \in \{250, 500, ..., 2500\}$. In addition, USF2Oracle and USB2Oracle are provided where the true skeleton is used while the ordering is inferred via Algorithms 1 and 2, respectively.

Figure 4 shows that both USF2 and USB2 more accurately recover the true directed edges as sample size increases. In addition, the performance of USF2 becomes perfect. This result supports the main results that non-linear ANMs with homogeneous and heterogeneous error variances can be identifiable, and our algorithms are consistent as long as the required conditions are met.

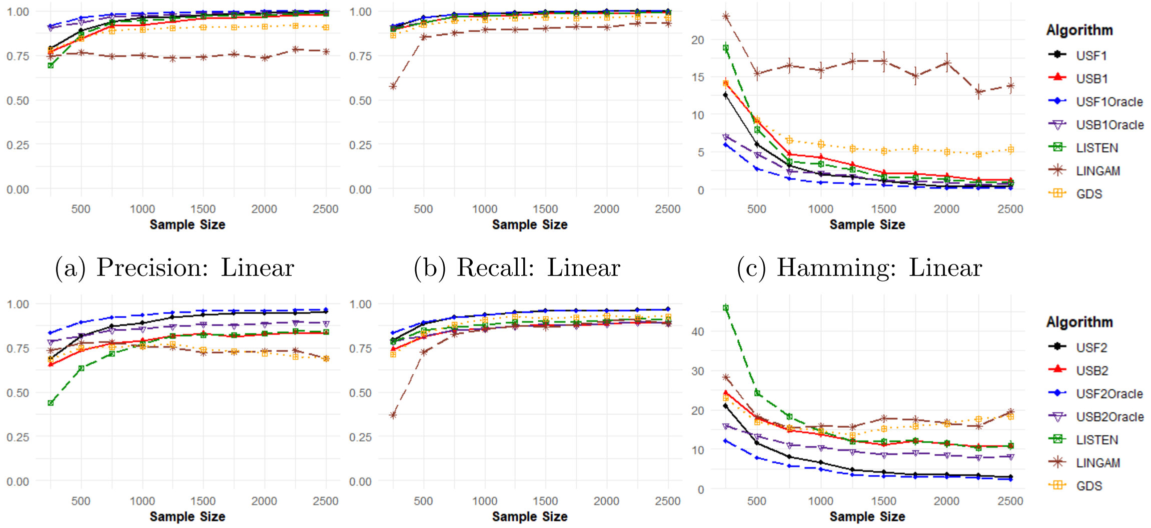
It must be pointed out that, given the same sample size, USB2(Ordering) is way worse than USF2(Ordering), and its the hamming is far way from 0. This represents that the forward stepwise selection condition is milder than the backward stepwise selection condition in our non-linear settings. Hence, it makes sense that USF2 performs significantly better than USB2 as well as the comparison algorithms. More precisely, the GDS, LISTEN, and LINGAM algorithms cannot recover polynomial SEMs because they are designed only for learning linear SEMs. Lastly, we emphasize that this result does not imply that the forward selection condition is strictly weaker than the backward selection condition in a general setting.

### 5.4. Random ANMs with Gaussian and Non-Gaussian Errors

This section verifies one of our main results that non-Gaussian ANMs can be identifiable where non-Gaussian error distributions are allowed. Hence, 100 sets of samples were generated under the same procedure specified in Sections 5.1 and 5.3, except that error distributions were sequentially uniform, $U(-1, 1)$, Gaussian $N(0, 1/3)$, and a half of t-distribution with 10 degree of freedom.

Then, our algorithms and the comparison methods are evaluated by varying sample size $n \in \{250, 500, ..., 2500\}$ as seen in Figure 5. Step (1) of Algorithm 3 was implemented using a linear model for linear SEM and a quadratic model for polynomial SEMs. In addition, Step (2) of Algorithm 3 was implemented using a permutation test with mutual information, since non-Gaussian error distributions are considered. This procedure is clearly not an exact test for general continuous distributions, but finding the proper conditional independence test and its sample complexity is also left to future study.

The simulation results in Figure 5 are analogous to the results for Gaussian linear and polynomial SEMs in Sections 5.2, and 5.3. Our algorithms are estimating the true directed edges better as sample size is increasing. In addition, the precision and recall of USF1 and



(a) Precision: Linear (b) Recall: Linear (c) Hamming: Linear

(d) Precision: Quadratic (e) Recall: Quadratic (f) Hamming: Quadratic

Figure 5: Comparison of the proposed algorithms (USF1, USB1, USF2, USB2), the proposed algorithms with true skeletons (USF1Oracle, USB1Oracle, USF2Oracle, USB2Oracle), the GDS, LISTEN, and LINGAM algorithms in terms of average precision, recall, and Hamming distance for recovering 20-node linear and quadratic SEMs with heterogeneous uniform, normal, and t error distributions.

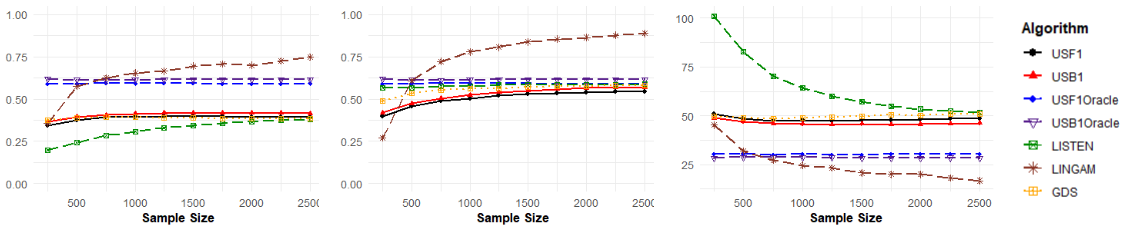
USF2 converge to 1, and their hamming distances also converge to 0. Hence, this confirms that ANMs with non-Gaussian error distributions can be identifiable, regardless of error distributions. In addition, like other simulation settings, USF1 and USF2 perform better than the comparison methods in terms of recovering DAGs on average.

So far, the simulation results have heuristically confirmed that Algorithm 3 can recover the proposed identifiable ANMs as long as the required conditions are satisfied. In other words, Algorithm 3 fail to recover a model when the proposed identifiability condition is violated. To emphasize this point, 100 sets of samples were again generated under the same procedure specified in Sections 5.1, except that error distributions were sequentially uniform, $U(-3, 3)$, Gaussian, $N(0, 0.25)$, and t-distribution with 10 degree of freedom. In this setting where the error variances are sequentially 3, 0.25, and 1.25, the proposed identifiability condition is clearly not satisfied.

As we can see in Figure 6, neither USF1Oracle and USB1Oracle can estimate the true orderings. And hence, it is expected that USF1, USB1, and LISTEN fail to recover the true graphs. However, it is worth noting that the LINGAM algorithm is for learning non-Gaussian linear SEMs, and hence, it shows a better performance.

## 6. Real Multivariate Data: Mathematics Marks

We applied our algorithms and the comparison GES, GDS, LISTEN, and LINGAM algorithms to real multivariate Gaussian data involving students mathematics scores. More precisely, the variables are the examination marks for 88 students from five different sub-



(a) Precision (b) Recall (c) Hamming Distance

Figure 6: Comparison of the proposed algorithms (USF1, USB1), the proposed algorithms with true skeletons (USF1Oracle, USB1Oracle), the GDS, LISTEN, and LINGAM algorithms in terms of average precision, recall, and Hamming distance for recovering 20-node linear SEMs with heterogeneous uniform, normal and students t error distributions.

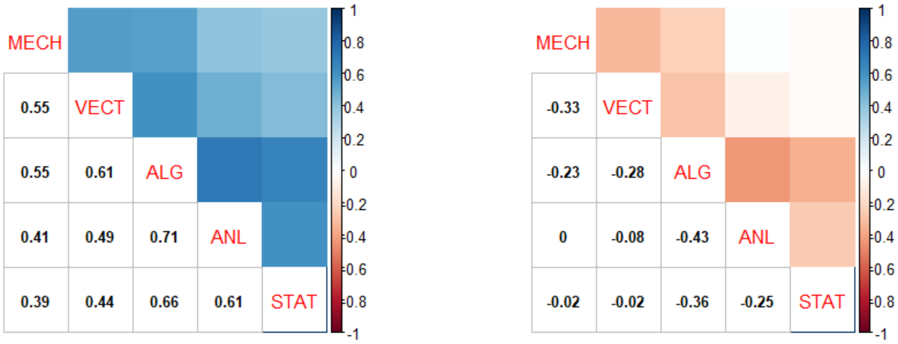
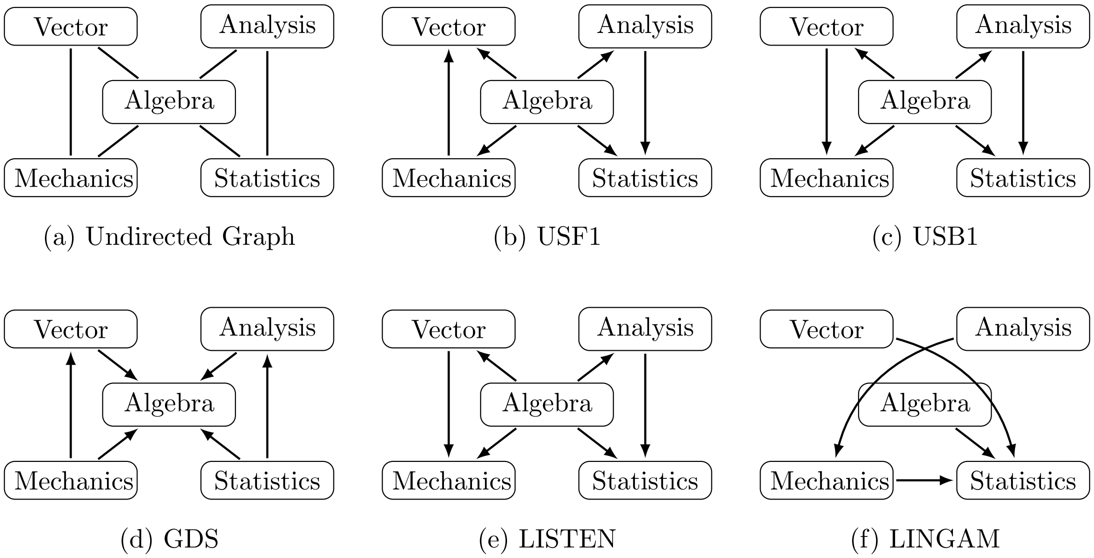


Figure 7: Correlation and partial correlation plots for students' examination scores.

jects: mechanics, vectors, algebra, analysis, and statistics. All are measured on the same scale (from 0 to 100). This dataset is provided in the bnlearn R package (Scutari, 2009).

As we can see on the left side of Figure 7, the correlation plot shows that all five variables are positively correlated. This means that students who do well in one subject are more likely to do well in the others. To see the conditional independence relationships, the partial correlation plot is provided on the right side of Figure 7. As we can see, some components are very close to 0, for example, between (mechanics, analysis), (mechanics, statistics), and (vectors, statistics). This shows that some subjects are conditionally independent given other subjects. That makes sense, because all other subjects' scores are not necessary to explain the score of a subject.

The mathematics marks data were originally modeled using the Gaussian undirected graphical model in Edwards (2012). The estimated undirected graph is provided in Figure 8. Edwards (2012) claims that the Gaussian undirected graphical model successfully captures the conditional independence relationships, as shown in Figure 7 (right). The scores for analysis and statistics are conditionally independent of mechanics and vectors, given algebra. Hence, the graph shows that for prediction of the statistics scores, the marks for algebra and analysis are sufficient, and for prediction of the analysis scores, the scores for algebra



(a) Undirected Graph (b) USF1 (c) USB1

(d) GDS (e) LISTEN (f) LINGAM

Figure 8: The examination marks' undirected graph and the directed acyclic graph estimated by the proposed algorithms (USF1, USB1) and the comparison GDS, LISTEN, and LINGAM algorithms.

and statistics are sufficient. In addition for prediction of the scores for algebra, the scores for all the other subjects are required.

We believe that there exist directional relationships between subjects, because one course can be essential to another course. For example, algebra is an evidently central subject for all other subjects. In addition, knowledge of analysis and vectors is prerequisite for statistics and mechanics, respectively. This may follow a linear SEM with constraints on linear weights, that is, $X_j = \beta_{j0} + \sum_{k \in \mathrm{Pa(j)}} \beta_{jk} X_k + \epsilon_j$ where $X_j$ is the score of a subject $j$ and $\mathrm{Pa}(j)$ is a set of prerequisites for a subject $j$.

Clearly, the undirected graph does not provide these directional relationships. Hence, in order to recover these directional relationships, our algorithms, USF1 and USB1, are applied using ordinary linear regression and Fisher's z-transform of the partial correlation. The significance level was set to $\alpha = 1 - \Phi(n^{1/4}/2)$ as in Section 5.1, and the estimated directed graph is provided in Figure 8. For the LISTEN algorithm, we used the regularization parameters to 0.001 and the hard threshold parameter was set to 0.25 because these seem to recover the model better.

As we can see in Figure 8, the undirected graph and all the estimated graphs via DAG learning algorithms, except for LINGAM, have the same skeleton, which implies that theses algorithms recover all important links. Moreover, USB1 and LISTEN find all the directional relationships between the subjects. USF1 also recovers most of the explainable directed edges while it returns a reversed edge between vector and mechanics. However, the estimated edges from GDS are all reversed; LINGAM estimates an unexplainable skeleton and its directed edges are not well interpretable. That is mainly because GDS applies a greedy search, and LINGAM does not guarantee recovering the directed edges when the data follow a Gaussian distribution. Lastly, the GES algorithm cannot find any directed

| | USF1 | USB1 | GDS | LINSTEN | LINGAM | GES |
|---|---|---|---|---|---|---|
| log-likelihood | -1695.589 | -1695.589 | -1711.135 | -1762.506 | -1695.589 | -1695.589 |
| BIC | -1731.407 | -1731.407 | -1746.954 | -1793.848 | -1731.407 | -1731.407 |

Table 1: Log-likelihood and BIC scores of the proposed algorithms (USF1, USB1) and state-of-the-art GES, GDS, LISTEN, and LINGAM algorithms.

edges, because it can only find directed edges if at least one v-structures exists while the true graph does not contain any v-structures.

It is granted that the true graphs considered may not represent true directional relationships between subjects. Hence, we also compared the log-likelihood and BIC scores of the algorithms in Table 1. As we can see in Table 1, the USB1, USF1, LISTEN, and GES algorithms produce the same scores while the GDS and LINGAM algorithms have lower scores. We point out that, unlike the GES algorithm, the other algorithms are not focusing on maximizing the log-likelihood or BIC scores. However, our methods and LISTEN algorithm also maximize the scores, and hence, they might be more useful than the GES algorithm because they find a graph rather than the MEC. Therefore, we believe that our methods and LISTEN more reliably recovers the directional/functional relationships between the scores of the five courses.

## 7. Conclusion and Future Works

In this paper, we proposed a new identifiability condition for ANMs with heterogeneous error variances which is a generalization of the identifiability conditions for linear SEMs. Hence, we proved that ANMs can be identifiable without assuming faithfulness, linearity, and Gaussianity. We also proposed the US algorithm for learning identifiable ANMs, and provided its consistency for Gaussian linear SEMs. The various numerical experiments support our theoretical results, and empirically confirm that the US algorithm can capture the dependency of variables in identifiable non-linear and non-Gaussian ANMs.

Several topics remain for future works. Although the proposed identifiability condition is a general version of the conditions for linear SEMs, it could be very restrictive. Hence, it is an important problem of finding a proper test whether the forward or backward selection condition is satisfied. However, to the best of our knowledge, the conditions cannot be confirmed from data, and it should be investigated in the future. Furthermore, the statistical consistency of learning non-linear and non-Gaussian ANMs are not yet provided. We conjecture that the models can be learned in a consistent way, and one may be able to prove this.

## Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1C1B5085420). I thank all members of my research group, especially Sion Park and Yesool Kim, for helpful discussions.

## References

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2003.

Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.

David Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2012.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, pages 6457–6466, 2017.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752. ACM, 2009.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.

Gunwoong Park and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49(1):276–292, 2020.

Gunwoong Park and Hyewon Park. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 158–166. PMLR, 16–18 Apr 2019a.

Gunwoong Park and Sion Park. High-dimensional poisson structural equation model learning via $\ell_1$-regularized regression. *Journal of Machine Learning Research*, 20(95):1–41, 2019b.

Gunwoong Park and Garvesh Raskutti. Learning large-scale poisson dag models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, pages 631–639, 2015.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18(224): 1–44, 2018.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 597–604, 2010.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15 (1):2009–2053, 2014.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 491–498. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the ninth international workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers: Key West, FL, USA, 2003.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Y Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59, 2020.

Jiji Zhang and Peter Spirtes. The three faces of faithfulness. *Synthese*, 193(4):1011–1027, 2016.

Kun Zhang and Aapo Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–585. Springer, 2009a.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009b.

## Appendix A. Proof for Theorem 2

**Proof** Without loss of generality, assume that the true ordering $\pi = (\pi_1, ..., \pi_p)$ is unique. In addition, for ease of notation, define $X_{1:j} = \{X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_j}\}$ and $X_{1:0} = \emptyset$. Now, we prove the identifiability of ANMs using mathematical induction, as follows:

(A) Forward stepwise selection:

**Step (1)** By Condition (A) in Theorem 2, for any node $k \in V \setminus \{\pi_1\}$, we have,

$$\mathrm{Var}(X_{\pi_1}) = \sigma^2_{\pi_1} < \sigma^2_k + \mathrm{Var}(\mathbb{E}(X_k \mid X_{\mathrm{Pa}(k)})) = \mathrm{Var}(X_k).$$

Therefore, the first element of the ordering $\pi_1$ can be correctly identified.

**Step (m-1)** For the $(m-1)^{th}$ element of the ordering, assume that the first $m-1$ elements of the ordering and their parents are correctly recovered.

**Step (m)** Now, consider the $m^{th}$ element of the ordering and its parents. By Theorem 2, for $k \in \{\pi_{m+1}, \cdots, \pi_p\}$,

$$\begin{aligned} \mathbb{E}(\mathrm{Var}(X_{\pi_m} \mid X_{1:(m-1)})) &= \sigma^2_{\pi_m} < \sigma^2_k + \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_k \mid X_{\mathrm{Pa}(k)}) \mid X_{1:(m-1)})) \\ &= d\mathbb{E}(\mathrm{Var}(X_k \mid X_{1:(m-1)})). \end{aligned}$$

Hence, we can choose the true $m^{th}$ element of the ordering $\pi_m$.

In terms of the parents search, it is clear that conditional independence relationships are naturally encoded by the factorization (1) and imply causal minimality (see details in Spirtes 1995; Peters and Bühlmann 2014). In ANM models, causal minimality states that, for any node $j \in V$ and one of its parents $k \in \mathrm{Pa}(j)$,

$$\forall \;\; \mathrm{Pa}(j) \setminus \{k\} \subset S \subset \mathrm{Nd}(j) \setminus \{k\}, \qquad X_j \not\perp\!\!\!\perp X_k \mid X_S.$$

Therefore, we can choose the correct parents of $\pi_m$, that is, $\mathrm{Pa}(\pi_m) := \{k \in \{\pi_1, ..., \pi_{m-1}\} \mid X_k \not\perp\!\!\!\perp X_{\pi_m} \mid X_{\pi_1}, ..., X_{\pi_{m-1}} \setminus X_k\}$. By mathematical induction, this completes the proof.

(B) Backward stepwise selection:

**Step (1)** By Condition (B) in Theorem 2, for any node $\ell \in V \setminus \{\pi_p\}$, we have

$$\mathrm{Var}(X_{\pi_p} \mid X_{V \setminus \pi_p}) = \sigma^2_{\pi_p} > \sigma^2_\ell - \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_\ell \mid X_{V\setminus \pi_p}) \mid X_{\mathrm{Pa}(\ell)})) = \mathrm{Var}(X_\ell \mid X_{V \setminus \ell}).$$

Therefore, the last element of the ordering $\pi_p$ can be correctly identified. In addition, causal minimality implies that if a node $k$ is a parent of $\pi_p$, we have $X_j \not\perp\!\!\!\perp X_k \mid X_{V\setminus\{j,k\}}$ ; otherwise, $X_j$ and $X_k$ are conditionally independent. Hence, the parents of $\pi_p$ can also be recovered.

**Step (p-m)** For the $(m+1)^{th}$ element of the ordering, assume that the last $m+1$ elements of the ordering and their parents are correctly recovered.

**Step (p-m+1)** Now, we consider the $m^{th}$ element of the causal ordering and its parents. By Condition (B) in Assumption 2, for $\ell \in \{\pi_1, \pi_2, \cdots, \pi_{m-1}\}$,

$$\begin{aligned} \mathrm{Var}(X_{\pi_m} \mid X_{V \setminus \{\pi_m, ..., \pi_p\}}) = \sigma^2_{\pi_m} &> \sigma^2_\ell - \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_\ell \mid X_{V\setminus\{\ell, \pi_{m+1}, ..., \pi_p\}}) \mid X_{\mathrm{Pa}(\ell)})) \\ &= \mathrm{Var}(X_\ell \mid X_{V \setminus \{\ell, \pi_{m+1}, ..., \pi_p\}}). \end{aligned}$$

Hence, we can choose the true $m^{th}$ element of the ordering $\pi_m$. In terms of parents search, it can again be determined by conditional independence relationships. Therefore, we can choose the correct parents of $\pi_m$. By mathematical induction, this completes the proof.

■

## Appendix B. Proof for Theorem 4

**Proof** Again, without loss of generality, we assume that the true ordering $\pi = (\pi_1, ..., \pi_p)$ is unique. In addition for ease of notation, we define $X_{1:j} = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_j})$ and $X_{1:0} = \emptyset$.

(A) Forward stepwise selection: We note that the diagonal component of the inverse covariance matrix of $X$ is closely related to the conditional variance. Then, for any node $j \in \{\pi_j, ..., \pi_p\}$ and $S \subset \{\pi_1, \pi_2, ..., \pi_{j-1}\}$, we have,

$$\begin{aligned}[(\Sigma_{S\cup j, S\cup j})^{-1}]_{jj} &= \Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj} = \mathbb{E}(\mathrm{Var}(X_j \mid X_S)) \\ &= \sigma_j^2 + \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}) \mid X_S)).\end{aligned}$$

From Equation (5), we have

$$\Sigma = (I - B)^{-T}\Sigma_\epsilon(I - B)^{-1}.$$

In addition, using the Neumann power series, we obtain

$$(I - B)^{-1} = (I + B + B^2 + B^3 + ... + B^p).$$

Using the path interpretation, it is clear that for directed acyclic graphs, $B^k$ encodes the length $k$ path, and hence, matrix $B^p$ is the zero-matrix since there is no cycle.

Then, we have the covariance matrix using $(I - B)^{-1}$:

$$\Sigma = (I - B)^{-T}\Sigma_\epsilon(I - B)^{-1} = (I + A + A^2 + ... + A^p)\Sigma_\epsilon(I + B + B^2 + ... + B^p)$$

where $A$ is a transpose of $B$.

For ease of notation, we define

$$D^T = \Sigma_\epsilon^{1/2}(1 + B + B^2 + ... + B^p)$$

Then, using the path interpretation, the $(j,k)$-th element of $[D]_{jk}$ corresponds to the sum of influences, $\beta_{j\to k}$, along with all directed paths from $j$ to $k$ of products of coefficients $(\beta_{jk})$ along each path. In addition, matrix $D$ is a lower triangular matrix.

Then, $\Sigma_{S\cup j, S\cup j}$ is the product of a matrix that can be partitioned into four blocks, and can be inverted block-wise as follows: For any node $j \in \{\pi_{k+1}, ..., \pi_p\}$,

$$\Sigma_{S\cup j, S\cup j} = \begin{bmatrix} D_1 & \mathbf{0} \\ D_3 & D_4 \end{bmatrix} \begin{bmatrix} D_1 & \mathbf{0} \\ D_3 & D_4 \end{bmatrix}^T,$$

where

$$D_1 = [\Sigma_\epsilon^{1/2}]_{S,S} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \beta_{\pi_1\pi_2} & 1 & \cdots & 0 \\ \beta_{\pi_1\to\pi_3} & \beta_{\pi_2\pi_3} & \cdots & 0 \\ \cdots & & & \\ \beta_{\pi_1\to\pi_{j-1}} & \beta_{\pi_2\to\pi_{j-1}} & \cdots & 1 \end{bmatrix}, D_3^T = \begin{bmatrix} \beta_{\pi_1\to j}\sigma_{\pi_1} \\ \cdots \\ \beta_{\pi_{j-1}\to j}\sigma_{\pi_{j-1}} \end{bmatrix}, D_4^T = \begin{bmatrix} \beta_{\pi_j\to j}\sigma_{\pi_j} \\ \cdots \\ \beta_{\pi_k\to j}\sigma_{\pi_k} \\ \sigma_j \end{bmatrix}.$$

The $\mathbf{0} \in \mathbb{R}^{j-1\times j-1}$ is a zero matrix from the definition of the ordering, in that there is no direct path from $\pi_k \to ... \to \pi_j$ when $j < k$.

Then, the inversion of the parts of the covariance matrix is as follows:

$$[(\Sigma_{S\cup j,S\cup j})^{-1}]_{j,j} = \Big(D_3{D_3}^T + D_4{D_4}^T - D_3D_1(D_1D_1^T)^{-1}{D_1}^T{D_3}^T\Big)^{-1}$$

Since $D_1$ is a lower triangular matrix, we have

$$[(\Sigma_{S\cup j,S\cup j})^{-1}]_{j,j} = \big(D_3{D_3}^T + D_4{D_4}^T - D_3{D_3}^T\big)^{-1} = (D_4{D_4}^T)^{-1}.$$

It can be rewritten using the edge weight and error variances, as follows:

$$\mathrm{Var}(X_j \mid X_S) = (D_4{D_4}^T) = \sigma_j^2 + \sum_{k\in \mathrm{Pa}(j)\setminus S} \beta_{k\to j}^2\sigma_k^2,$$

Finally, for any $m \in \{1,2,...,p\}$, let $j = \pi_m$, $k \in \mathrm{De}(j)$, and $1:j = \{\pi_1,\pi_2,..,\pi_m\}$. Therefore, Assumption 2 (A) is equivalent to

$$\sigma_j^2 < \sigma_k^2 + \sum_{k'\in \mathrm{Pa}(k)\setminus 1:(j-1)} \beta_{k'\to k}^2\sigma_{k'}^2.$$

(B) Backward stepwise selection: Again, note that the diagonal component of the inverse covariance matrix of $X$ is as follows (see more details in Ghoshal and Honorio, 2018; Loh and Bühlmann, 2014):

$$\Omega_{jj} = \frac{1}{\sigma_j^2} + \sum_{\ell\in \mathrm{Ch}(j)} \frac{\beta_{j\ell}^2}{\sigma_\ell^2}.$$

Then, for any node $j \in V$ and $S = V \setminus j$, we have

$$(\Omega_{jj})^{-1} = \Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj} = \mathbb{E}(\mathrm{Var}(X_j \mid X_S)) = \sigma_j^2 - \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_j \mid X_S) \mid X_{\mathrm{Pa}(j)})).$$

Therefore, for any terminal node $j$ and non-terminal node in the ordering $\ell$,

$$\sigma_j^2 > \sigma_\ell^2 - \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_\ell \mid X_S) \mid X_{\mathrm{Pa}(\ell)})) \iff \Omega_{jj} < \Omega_{\ell\ell} \iff \frac{1}{\sigma_j^2} < \frac{1}{\sigma_\ell^2} + \sum_{k\in \mathrm{Ch}(\ell)} \frac{\beta_{\ell k}^2}{\sigma_k^2}$$

Since the remainder of the proof is exactly the same after eliminating the terminal node, we can omit it.

■

## Appendix C. Proof for Theorem 8

### Proof

Without loss of generality, assume that the true ordering is unique, and that $\pi = (\pi_1, ..., \pi_p) = (1, 2, ..., p)$ and $\pi_0 = \emptyset$. For ease of notation, define the marginal variance of $X_j$ as $\sigma^2(j)$, and the conditional variance of $X_j$ given $X_S$ as $\sigma^2(j, S)$. Lastly, let $\pi_{1:j} = (\pi_1, ..., \pi_j)$. Then, the probability that the ordering is correctly estimated from Algorithm 1 is

$$\begin{aligned} &P\left(\widehat{\pi} = \pi\right) \\ =&P\left(\widehat{\sigma}^2(1) < \min_{j=2,...,p} \widehat{\sigma}^2(j),\ \widehat{\sigma}^2(2, \pi_1) < \min_{j=3,...,p} \widehat{\sigma}^2(j, \pi_1), ...,\ \widehat{\sigma}^2(p-1, \pi_{1:p-2}) < \widehat{\sigma}^2(p, \pi_{1:p-2})\right) \\ =&P\left(\min_{j=1,...,p-1} \min_{k=j+1,...,p} \widehat{\sigma}^2(k, \pi_{1:j-1}) - \widehat{\sigma}^2(j, \pi_{1:j-1}) > 0\right). \end{aligned}$$

Since it can be decomposed in to the following three terms, we have

$$\begin{aligned} &P\left(\widehat{\pi} = \pi\right) \\ =&P\Bigg(\min_{\substack{j=1,...,p-1 \\ k=j+1,...,p}} \Big\{\Big(\sigma^2(k, \pi_{1:j-1}) - \sigma^2(j, \pi_{1:j-1})\Big) + \big(\sigma^2(j, \pi_{1:j-1}) - \widehat{\sigma}^2(j, \pi_{1:j-1})\big) \\ &\qquad\qquad - \big(\sigma^2(k, \pi_{1:j-1}) - \widehat{\sigma}^2(k, \pi_{1:j-1})\big) > 0\Big\}\Bigg) \\ \geq&P\left(\min_{\substack{j=1,...,p-1 \\ k=j+1,...,p}} \Big\{\Big(\sigma^2(k, \pi_{1:j-1}) - \sigma^2(j, \pi_{1:j-1})\Big)\Big\} > \tau_F \ , and \right. \\ &\qquad\qquad \left. \max_{\substack{j=1,...,p-1 \\ k=j,...,p}} \left|\sigma^2(k, \pi_{1:j-1}) - \widehat{\sigma}^2(k, \pi_{1:j-1})\right| < \frac{\tau_F}{2}\right). \end{aligned}$$

The first term in the above probability is always satisfied because $\sigma^2(k, \pi_{1:j-1}) - \sigma^2(j, \pi_{1:j-1}) > \tau_F$ from Assumption 7. Hence, the lower bound of the probability that the ordering is correctly estimated using our method is reduced to

$$P\left(\widehat{\pi} = \pi\right) \geq P\left(\max_{\substack{j=1,...,p-1 \\ k=j,...,p}} \left|\sigma^2(k, \pi_{1:j-1}) - \widehat{\sigma}^2(k, \pi_{1:j-1})\right| < \frac{\tau_F}{2}\right).$$

In a linear SEM setting, a conditional variance of $X_j$ given $X_S$ can be expressed as $\Sigma_{jj} - \Sigma_{j,S}\Sigma_{S,S}^{-1}\Sigma_{S,j}$. Then, we have,

$$\begin{aligned} |\widehat{\text{Var}}(X_j \mid X_S) - \text{Var}(X_j \mid X_S)| &= \left|\left(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{j,S}\widehat{\Sigma}_{S,S}^{-1}\widehat{\Sigma}_{S,j}\right) - \left(\Sigma_{jj} - \Sigma_{j,S}\Sigma_{S,S}^{-1}\Sigma_{S,j}\right)\right| \\ &\leq \underbrace{\left|\left(\widehat{\Sigma}_{jj} - \Sigma_{jj}\right)\right|}_{A} + \underbrace{\left|\ \left(\widehat{\Sigma}_{j,S}\widehat{\Sigma}_{S,S}^{-1}\widehat{\Sigma}_{S,j}\right) - \left(\Sigma_{j,S}\Sigma_{S,S}^{-1}\Sigma_{S,j}\right)\right|}_{B}. \end{aligned}$$

The first term, marked $A$, is bounded directly from Lemma 1 of Ravikumar et al. (2011):

$$\mathbb{P}\left(\max_{j,k\in V}\left|(\widehat{\Sigma}_{jk}-\Sigma_{jk})\right|\geq\frac{1}{\sqrt{\log n}}\right)\leq 4\cdot\exp\left(\frac{-n}{C_1\log n}\right),$$

where $C_1=1600\max_j(\Sigma_{jj})^2$.

The second term, marked $B$, is also bounded by the following three terms:

$$\begin{aligned}&\left|\left(\widehat{\Sigma}_{j,S}\widehat{\Sigma}_{S,S}^{-1}\widehat{\Sigma}_{S,j}\right)-\left(\Sigma_{j,S}\Sigma_{S,S}^{-1}\Sigma_{S,j}\right)\right|\\&\leq\left|\widehat{\Sigma}_{j,S}(\widehat{\Sigma}_{S,S}^{-1}-\Sigma_{S,S}^{-1})\widehat{\Sigma}_{S,j}\right|+\left|\widehat{\Sigma}_{j,S}\Sigma_{S,S}^{-1}(\widehat{\Sigma}_{S,j}-\Sigma_{S,j})\right|+\left|(\widehat{\Sigma}_{j,S}-\Sigma_{j,S})\Sigma_{S,S}^{-1}\Sigma_{S,j}\right|\\&\leq\|\widehat{\Sigma}_{j,S}\|_2\Lambda_{\max}(\widehat{\Sigma}_{S,S}^{-1}-\Sigma_{S,S}^{-1})+\Lambda_{\max}(\Sigma_{S,S}^{-1})\|\widehat{\Sigma}_{j,S}-\Sigma_{j,S}\|_2(\|\widehat{\Sigma}_{j,S}\|_2+\|\Sigma_{S,j}\|_2),\end{aligned}$$

where $\Lambda_{\max}(A)$ is a maximum eigenvalue of $A$.

This is bounded directly from Lemma 29 of Loh and Bühlmann (2014) for a given $j\in V$ and $S\subset V$; for sufficiently large $n$ such that $2\max_{j\in V}\sigma_j^2\Lambda_{\max}(\Sigma^{-1})\leq\sqrt{\log n}$, we have

$$\mathbb{P}\left(\Lambda_{\max}(\widehat{\Sigma}_{S,S}-\Sigma_{S,S})\right)\geq\frac{1}{\sqrt{\log n}}\right)\leq 2\mathrm{exp}\left(-C_2\frac{n}{\log n}\right)$$

and

$$\mathbb{P}\left(\Lambda_{\max}(\widehat{\Sigma}_{S,S}^{-1}-\Sigma_{S,S}^{-1})\right)\geq\frac{\Lambda_{\max}(\Sigma^{-1})}{\sqrt{\log n}}\right)\leq 2\mathrm{exp}\left(-C_3\frac{n}{\log n}\right).$$

In addition, we have

$$\|\widehat{\Sigma}_{j,S}-\Sigma_{j,S}\|_2\leq\Lambda_{\max}(\widehat{\Sigma}_{S',S'}-\Sigma_{S',S'}),$$

where $S'=\{j\}\cup S$. Furthermore, we also have,

$$\|\widehat{\Sigma}_{j,S}\|_2\leq\|\Sigma_{j,S}\|_2+\Lambda_{\max}(\widehat{\Sigma}_{S',S'}-\Sigma_{S',S'})$$

Hence, for a sufficiently large $n$, there exist positive constants $M_1$ and $C_4$ such that

$$\begin{aligned}&|\widehat{\mathrm{Var}}(X_j\mid X_S)-\mathrm{Var}(X_j\mid X_S)|\\&\leq\frac{1}{\sqrt{\log n}}+\left(\|\Sigma_{j,S}\|_2+\frac{1}{\sqrt{\log n}}\right)\frac{\Lambda_{\max}(\Sigma^{-1})}{\sqrt{\log n}}+\frac{\Lambda_{\max}(\Sigma^{-1})}{\sqrt{\log n}}\left(2\|\Sigma_{j,S}\|_2+\frac{1}{\sqrt{\log n}}\right)\\&\leq\frac{M_1}{\sqrt{\log(n)}}\leq\frac{\tau_F}{2}\end{aligned}$$

with a probability of at least $1-\exp(-C_4n/\log n)$. Therefore, for fixed node size $p$, the proof for Algorithm 1 is complete.

$$P\left(\widehat{\pi}=\pi\right)\geq 1-\frac{p(p-1)}{2}\mathrm{exp}\left(-C_4\frac{n}{\log n}\right).$$

Now, we provide the proof for Algorithm 2. In a similar method, the probability that the ordering is correctly estimated from our backward stepwise selection method can be written as

$$
\begin{aligned}
&P\left(\widehat{\pi}=\pi\right) \\
&= P\Big(\widehat{\sigma}^2(p,\pi_{1:p-1}) > \min_{j=1,...,p-1} \widehat{\sigma}^2(j, V\setminus j),\ \widehat{\sigma}^2(p-1,\pi_{1:p-2}) > \min_{j=1,...,p-2} \widehat{\sigma}^2(j,\pi_{1:p-1}\setminus j) \\
&\qquad\qquad\qquad\qquad ,...,\ \widehat{\sigma}^2(2,1) > \widehat{\sigma}^2(1,2)\Big) \\
&= P\Big( \min_{j=2,...,p} \min_{k=1,...,j-1} \widehat{\sigma}^2(j,\pi_{1:j}\setminus j) - \widehat{\sigma}^2(k,\pi_{1:j}\setminus k) > 0\Big)
\end{aligned}
$$

Again, since it can be decomposed in to the following three terms, we have

$$
\begin{aligned}
&P\left(\widehat{\pi}=\pi\right) \\
&= P\Big( \min_{\substack{j=2,...,p\\ k=1,...,j-1}} \Big\{ \Big(\sigma^2(j,\pi_{1:j}\setminus j) - \sigma^2(k,\pi_{1:j}\setminus k)\Big) - \big(\sigma^2(j,\pi_{1:j}\setminus j) - \widehat{\sigma}^2(j,\pi_{1:j}\setminus j)\big) \\
&\qquad\qquad\qquad\qquad + \big(\sigma^2(k,\pi_{1:j}\setminus k) - \widehat{\sigma}^2(k,\pi_{1:j}\setminus k)\big) > 0 \Big\}\Big) \\
&\geq P\Big( \min_{\substack{j=2,...,p\\ k=1,...,j-1}} \Big\{ \Big(\sigma^2(j,\pi_{1:j}\setminus j) - \sigma^2(k,\pi_{1:j}\setminus k)\Big)\Big\} > \tau_B, \text{ and} \\
&\qquad\qquad\qquad\qquad \max_{\substack{j=2,...,p\\ k=1,...,j}} \big|\sigma^2(k,\pi_{1:j}\setminus k) - \widehat{\sigma}^2(k,\pi_{1:j}\setminus k)\big| < \frac{\tau_B}{2}\Big).
\end{aligned}
$$

Similar to the proof for Algorithm 1, the first term in the above probability is always satisfied, because $\sigma^2(j,\pi_{1:j}\setminus j) - \sigma^2(k,\pi_{1:j}\setminus k) > \tau_B$ from Condition (B) in Assumption 7. Hence, the lower bound of the probability that the ordering is correctly estimated via Algorithm 2 is

$$
P\left(\widehat{\pi}=\pi\right) \geq P\left( \max_{\substack{j=2,...,p\\ k=1,...,j}} \big|\sigma^2(j,\pi_k) - \widehat{\sigma}^2(j,\pi_k)\big| < \frac{\tau_B}{2}\right).
$$

Since the remaining proof is analogous to the above proof for Algorithm 1, we can omit the detail of the proof. Hence, we prove that for a fixed $p$ and a sufficiently large $n$, there exists a positive constant $C_4$ such that

$$
P\left(\widehat{\pi}=\pi\right) \geq 1 - \frac{p(p-1)}{2}\exp\left(-C_4\frac{n}{\log n}\right).
$$

■

## Appendix D. Proof for Theorem 9

### Proof

Without loss of generality, we suppose that the true ordering is $\pi = (1, 2, ..., p)$. We apply partial correlations to test the conditional independencies between variables. The sample partial correlation $\hat{\rho}_{j,k,S}$ can be calculated via linear regression, the inversion of the parts of the covariance matrix, or recursively by using the following identity: for some $s \in S$,

$$\hat{\rho}_{j,k,S} = \frac{\hat{\rho}_{j,k,S\setminus s} - \hat{\rho}_{j,s,S\setminus s}\hat{\rho}_{k,s,S\setminus s}}{\sqrt{(1-\hat{\rho}^2_{j,s,S\setminus s})(1-\hat{\rho}^2_{k,s,S\setminus s})}}.$$

With the partial correlations, we apply Fishers z-transform for the conditional independence tests:

$$Z_{j,k,S} = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}_{j,k,S}}{1-\widehat{\rho}_{j,k,S}}\right).$$

As seen in Kalisch and Bühlmann (2007), our proof is based on the following lemma:

**Lemma 11 (Lemma3 in Kalisch and Bühlmann, 2007)** *Suppose that Assumption 6 is satisfied. Then, for any* $\gamma > 0$,

$$\sup_{j,k,S} \mathbb{P}(|Z_{j,k,S} - z_{j,k,S}| > \gamma) \leq O(n-d)\left( exp\left\{ (n-p-4)\log\left(\frac{4-(\gamma/L)^2}{4+(\gamma/L)^2}\right)\right\} + exp\left\{ -C_2(n-p)\right\}\right),$$

*for some constant* $0 < C_2 < \infty$ *and* $L = 1/(1-(1+M)^2/4)$.

Consider a pair of nodes $(j,k)$ where $j < k$, which means $j$ is not a descendant of $k$: $j \notin \mathrm{De}(k)$. In addition, we consider a conditioning set $S$ such that $\mathrm{Pa}(j) \subset S \subset V \setminus \mathrm{De}(j)$. Lastly, let $E_{j,k,S}$ be an error event that consists of type I and II errors where $E_{j,k,S} = E^I_{j,k,S} \cup E^{II}_{j,k,S}$.

Then, each type of error is as follows:

$$\text{type I Error}E^I_{j,k,S} : \sqrt{n-|S|-3}\,|Z_{j,k,S}| > \Phi^{-1}(1-\alpha/2) \text{ when } z_{j,k,S} = 0,$$
$$\text{type II Error}E^{II}_{j,k,S} : \sqrt{n-|S|-3}\,|Z_{j,k,S}| \leq \Phi^{-1}(1-\alpha/2) \text{ when } z_{j,k,S} \neq 0,$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

Let significance level $\alpha(n) = 2(1-\Phi(\sqrt{n}\cdot\kappa(n)/2))$ where $\kappa(n)$ is as expressed in Assumption 6. By Lemma 11, we have

$$\sup_{j,k,S} \mathbb{P}(E^I_{j,k,S}) = \sup_{j,k,S} \mathbb{P}\left(|Z_{j,k,S} - z_{j,k,S}| > \sqrt{\frac{n}{n-|S|-3}}\frac{\kappa(n)}{2}\right)$$
$$\leq O(n-p)\mathrm{exp}\left(-C_3(n-p)\kappa(n)^2\right),$$

for some positive constant $C_3 > 0$ (see more details in Lemma 4 of Kalisch and Bühlmann, 2007).

In addition, we have

$$\begin{aligned}\sup_{j,k,S} \mathbb{P}(E^{II}_{j,k,S}) &= \sup_{j,k,S} \mathbb{P}\left(|Z_{j,k,S}| < \sqrt{\frac{n}{n-|S|-3}}\frac{\kappa(n)}{2}\right) \\ &= \sup_{j,k,S} \mathbb{P}\left(|Z_{j,k,S} - z_{j,k,S}| > (1 - \sqrt{\frac{n}{4(n-|S|-3)}})\kappa(n)\right) \\ &\leq O(n-p)\text{exp}(-C_4(n-p)\kappa(n)^2),\end{aligned}$$

for some positive constant $C_4 > 0$ using Lemma 11 (see more details in Lemma 4 of Kalisch and Bühlmann, 2007).

Therefore, using the union bound, we have

$$\mathbb{P}(E_{j,k,S}) \leq O(n-p)\text{exp}(-C_5(n-p)\kappa(n)^2).$$

for a positive constant $C_5 > 0$.

Given the true ordering, there are $p(p-1)/2$ hypothesis tests. Hence,

$$\mathbb{P}(\text{ an error occurs in our algorithm}) \leq O(p^2(n-p))\text{exp}\left(-C_5(n-p)\kappa(n)^2\right).$$

Therefore, for a sufficiently large sample size $n$, and for a fixed node size $p$, our algorithm recovers the true graph with a high probability. This completes the proof. ■