Identifiability of Gaussian Structural Equation Models with Homogeneous and Heterogeneous Error Variances

Gunwoong Park[1]

[1] Department of Statistics, University of Seoul

Abstract

In this work, we consider the identifiability assumption of Gaussian structural equation models (SEMs) in which each variable is determined by a linear function of its parents plus normally distributed error. It has been shown that linear Gaussian structural equation models are fully identifiable if all error variances are the same or known. Hence, this work proves the identifiability of Gaussian SEMs with both homogeneous and heterogeneous unknown error variances. Our new identifiability assumption exploits not only error variances, but edge weights; hence, it is strictly milder than prior work on the identifiability result. We further provide a structure learning algorithm that is statistically consistent and computationally feasible, based on our new assumption. The proposed algorithm assumes that all relevant variables are observed, while it does not assume causal minimality and faithfulness. We verify our theoretical findings through simulations, and compare our algorithm to state-of-the-art PC, GES and GDS algorithms.

Keywords: Causal inference, identifiability, directed acyclic graphical model, Bayesian network, Gaussian structural equation modeling, multiple regressions

1 Introduction

Learning the causal structure of a set of random variables from joint distribution is an important problem in many areas (Kephart and White 1991; Friedman et al. 2000; Doya 2007; Peters and Bühlmann 2014). This problem becomes more crucial when the causal graph is of interest but interventional experiments are impossible. However, learning causal graphical models from only observational data is a notoriously difficult problem due to non-identifiability. Hence, a number of prior works have addressed the question of identifiability for different classes of joint distribution by placing further restrictions on distribution $P(G)$. Spirtes et al. (2000), Chickering (2003), Tsamardinos and Aliferis (2003), Zhang and Spirtes (2016) and many other works show that directed acyclic graphical (DAG) models are recoverable up to the Markov equivalence class (MEC) under the faithfulness or related assumptions. However, since many MECs contain more than one graph, a true causal graph cannot be determined.

Recent works prove a number of fully identifiable classes of DAG models by placing a different type of restrictions on $P(G)$: (i) Shimizu et al. (2006) shows that linear non-Gaussian models where each variable is determined by a linear function of its parents plus an independent non-Gaussian error term are identifiable; (ii) Hoyer et al. (2009); Mooij et al. (2009); Peters et al. (2012) relax the assumption of linearity, and

prove the identifiability of nonlinear additive noise models where each variable is determined by a nonlinear function of its parents and an error term; (iii) Peters and Bühlmann (2014) proves that Gaussian linear structural equation models with equal or known error variances are identifiable; and (iv) Park and Raskutti (2018); Park (2018) prove the identifiability of DAG models where the variance of the conditional distribution of each node given its parents is a non-concave function of the mean.

In this article, we prove the identifiability of a new class of DAG models: Gaussian linear structural equation models with unknown error variances that can be different. Our approach exploits an uncertainty level of conditional distribution by considering both error variances and edge weights. We show that the new identifiability assumption is strictly milder than the equal error variance assumption for the Gaussian linear structural equation models in Peters and Bühlmann (2014).

In addition, we develop a statistically consistent and computationally feasible algorithm to recover a graph based on our new identifiability condition. We compare our algorithm against state-of-the-art PC Spirtes et al. (2000) greedy equivalence search (GES) Chickering (2003), and greedy DAG search (GDS) Peters and Bühlmann (2014) algorithms in Section 4. Our algorithm performs better than the comparisons because our algorithm is not a heuristic search, but exploits a relaxed identifiability condition. Lastly, we emphasize that the new condition enables the proposed algorithm to be a polynomial-time complete search, and hence, it can learn large-scale graphs.

The remainder of this paper is structured as follows. Section 2 summarizes the necessary notations and problem settings, discusses Gaussian SEMs, and proves their identifiability. In Section 3, we introduce a practical graph-learning algorithm based on our theoretical findings. Section 4 provides an evaluation of our algorithm against other state-of-the-art DAG learning algorithms when recovering the graphs.

2 Gaussian Structural Equation Models and Identifiability

We first introduce some necessary notations and definitions for Gaussian structural equation models (SEMs) and directed acyclic graphical (DAG) models. Then, we give a detailed description of the previous work on the identifiability of Gaussian SEMs in Peters et al. (2012). Lastly, we propose a new identifiability condition.

2.1 Problem Set-up and Notations

A DAG $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \cdots, p\}$ and a set of directed edges $E \in V \times V$ 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 $\text{Pa}(k)$ 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 $\text{De}(k)$ denotes the set of all descendants of node k. The *non-descendants* of node k are $\text{Nd}(k) := V \setminus (\{k\} \cup \text{De}(k))$. 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.

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 $\mathcal{X}(S) := \times_{j \in S} \mathcal{X}_j$. 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. Then, a DAG model has the following factorization Lauritzen (1996):

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

where $P(X_j \mid X_{\text{Pa}(j)})$ is the conditional distribution of variable X_j given its parents $X_{\text{Pa}(j)}$.

Throughout the paper, we assume causal sufficiency that all variables have been observed. Causal sufficiency is assumed in many DAG model learning methods including Gaussian SEMs with identical errors in Peters and Bühlmann (2014). In addition, although learning a DAG model is deeply involved with causal inference, we present the main statement without using causal terminology.

2.2 Gaussian Structural Equation Models

The Gaussian structural equation model (SEM) we consider is a special case of Gaussian DAG models where the joint distribution is defined by the following linear structural equations:

$$X_j = \beta_{j0} + \sum_{k \in \text{Pa}(j)} \beta_{jk} X_k + \epsilon_j, \qquad \forall j \in V \tag{2}$$

where $(\epsilon_j)_{j \in V}$ are independent, but not identical Gaussian distributions, $N(0, \sigma_j^2)$.

It can be restated in the following matrix form:

$$(X_1, X_2, ..., X_p)^T = B_0 + B(X_1, ..., X_p)^T + (\epsilon_1, ..., \epsilon_p)^T \tag{3}$$

where $B_0 \in \mathbb{R}^p$ is an intercept vector, and $B \in \mathbb{R}^{p \times p}$ is an edge weight matrix with each element $[B]_{jk} = \beta_{jk}$, in which β_{jk} is the weight of an edge from X_j to X_k. Furthermore, $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_p)^T \sim N(\mathbf{0}_p, \Sigma_\epsilon)$ where $\mathbf{0}_p = (0, 0, ..., 0)^T \in \mathbb{R}^p$, and Σ_ϵ is a diagonal matrix with unknown variances $\sigma_1^2, \sigma_2^2, ..., \sigma_p^2$.

The edge weight matrix B encodes the structure under the *non-zero edge weights condition* where β_{jk} is non-zero if $k \in \text{Pa}(j)$; otherwise, $\beta_{jk} = 0$, as in other linear structural equation models (see details in Spirtes 1995; Peters and Bühlmann 2014). It is a natural condition that is in accordance with the intuitive understanding of causal relationships among variables. In our linear structural equation settings, Theorem 1 to 3 in Pearl (2014) and Lemma 4 in Peters and Bühlmann (2014) prove that the condition of the edge weights (β_{jk}) implies the widely held Markov and *causal minimality* conditions in many causal DAG models learning approaches (see e.g. Pearl 2014; Spirtes et al. 2000; Peters et al. 2012). Causal minimality means

that a joint distribution is not Markov with respect to a strict sub-graph of the true graph. In our settings, it means the following for any node $j \in V$ and one of its parents $k \in \text{Pa}(j)$:

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

Hence, causal minimality is a weak form of faithfulness Spirtes et al. (2000). As we discussed, faithfulness is commonly assumed for learning the Markov equivalence graph, such as in the PC Spirtes et al. (2000), the GES Chickering (2003), and the the max-min hill-climbing Tsamardinos et al. (2006) algorithms. However, in practice, it cannot be tested, and might be very restrictive in finite sample settings Uhler et al. (2013).

Without loss of generality, we assume that $\mathbb{E}(X_j) = 0$ for all $j \in V$. Then, the distribution of the Gaussian SEM in Equation (3) is as follows:

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

where $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix, and Σ_ϵ is a covariance matrix for errors ϵ. Then, its density 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 restated as:

$$f_G(x_1, x_2, ..., x_p; \Theta) = \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\Big(-\frac{1}{2}(x_1, ..., x_p)\Theta(x_1, ..., x_p)^T \Big). \tag{4}$$

As discussed, Peters and Bühlmann (2014) shows that Gaussian SEMs are identifiable under the non-zero edge weights and the same error variances assumptions. In other words, if the data are generated by a Gaussian SEM with different unknown error variances, it is not guaranteed to find the correct graph. The assumption of the exact same error variances might be reasonable for applications with variables from a similar domain, but it is still unrealistic for real-world data. Therefore, the main focus of this paper is to propose a strictly milder identifiability condition, which allows heterogeneous error variances, by utilizing not only the scale of error variances but that of edge weights (β_{jk}). We discuss the details of the new identifiability assumption in the next section.

2.3 Identifiability

In this section, we prove that how Gaussian DAG models with unknown homogeneous error variances are identifiable. To provide intuition, we explain how Gaussian SEMs are identifiable from only the distribution using bivariate Gaussian SEMs illustrated in Fig. 1: $G_1 : X_1 \sim N(0, \sigma_1^2)$ and $X_2 \mid X_1 \sim N(\beta_1 X_1, \sigma_2^2)$, $G_2 : X_2 \sim N(0, \sigma_2^2)$ and $X_1 \mid X_2 \sim N(\beta_2 X_2, \sigma_1^2)$, and $G_3 : X_1 \sim N(0, \sigma_1^2)$ and $X_2 \sim N(0, \sigma_2^2)$, where X_1 and X_2 are independent.

Now, we show how to determine if the underlying graph is either G_1, G_2, or G_3. For G_1, if the error variances ratio satisfies $\sigma_2^2/\sigma_1^2 > (1 - \beta_1^2)$, we can determine the components of the ordering from the first using marginal variances

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

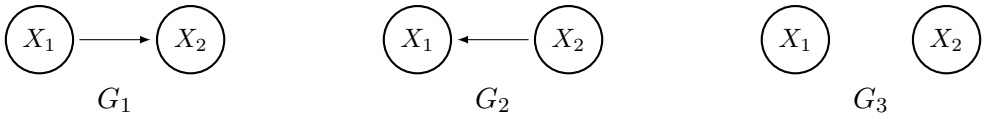


Figure 1: Bivariate directed acyclic graphs of G_1, G_2, and G_3

In addition, we can determine the ordering from the last using conditional variances

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

The error variance ratio condition $\sigma_2^2/\sigma_1^2 > (1 - \beta_1^2)$ is equivalent to the identifiability assumption in Theorem **??**. In addition, it holds under the same error variances and non-zero parameter β_1, which is the identifiability condition in Peters and Bühlmann (2014).

In the same manner, G_2 satisfies $\text{Var}(X_1) > \text{Var}(X_2)$ and $\text{Var}(X_1 \mid X_2) < \text{Var}(X_2 \mid X_1)$ as long as $\sigma_1^2/\sigma_2^2 > (1 - \beta_2^2)$, and hence, we can choose the true ordering $(2, 1)$. Lastly for G_3, there is no guarantee as to which marginal or conditional variance is bigger. However, either choice of ordering is fine since both $(1, 2)$ and $(2, 1)$ are correct orderings of G_3. Therefore, we can recover the orderings of G_1, G_2, and G_3 by testing which marginal or conditional variance is bigger.

Finding the skeleton procedure can be performed using the dependence relationships between variables. For G_1 and G_2, X_1 and X_2 are dependent under the minimality condition that is implied by the non-zero edge weights assumption. Combined with the ordering, we can distinguish G_1 and G_2. For G_3, X_1 and X_2 are independent under the Markov condition, and therefore, we can recover the graph.

Now, we extend this to general p-variate Gaussian SEMs with unknown error variances. The key idea to extending model identifiability from the bivariate to the multivariate involves the comparisons of the (conditional) node variances.

Theorem 2.1 (Identifiability)**.** *Let $P(X)$ be generated from a Gaussian SEM (4) with directed acyclic graph G. Suppose that Π_G is a true set of orderings for graph G. For any $m \in \{1, 2, ..., p\}$, let $j = \pi_m$, $k \in V \setminus Nd(j)$, $\ell \in An(j)$, and $1 : j = \{\pi_1, \pi_2, .., \pi_m\}$. Then, the DAG G is uniquely identifiable, if there exists $\pi \in \Pi_G$ satisfying either of the two following conditions*

(A) Forward Selection: $\sigma_j^2 < \sigma_k^2 + \mathbb{E}(Var(\mathbb{E}(X_k \mid X_{Pa(k)}) \mid X_{1:(j-1)}))$, or

(B) Backward Elimination: $\sigma_j^2 > \sigma_\ell^2 + \mathbb{E}(Var(\mathbb{E}(X_\ell \mid X_{1:(j-1)}) \mid X_{Pa(\ell)}))$,

We provide the detailed proof in Supplementary. Theorem 2.1 claims that Gaussian SEMs are identifiable if either the uncertainty level of a node j is smaller than that of its descendant, $\text{De}(j)$, given the non-descendant, $\text{Nd}(j)$ or the uncertainty level of a node j is bigger than that of its ancestor, $\text{An}(j)$, given the union of its parents and j, $\text{Pa}(j) \cup j$. The former condition can be understood that the noise of X_j is overestimated due to lack of parents. And, the latter can be understood that the noise of X_j is underestimated due to the full of parents plus addition of child that explains the error of X_j. As we will show later

in Algorithm **??**, condition (A) is applied to select a initial node, and condition (B) is used to determine a terminal node. Hence, the causal order can be identified.

Compared to the previous identifiability result for Gaussian linear SEMs in Theorem **??**, we can see the following result:

Lemma 2.2. *The identifiability condition (B) in Theorem 2.1 is equivalent to the condition in Theorem* **??**.

The proof is provided in Supplementary. Lemma 2.2 claims that our new identifiability condition is strictly weaker than the prior assumption because of a new condition (A) in Theorem 2.1.

Now we provide some special sufficient conditions for each assumption, and hence, uniquely recovers a Gaussian linear SEM:

Proposition 2.3. *The either of the following three conditions are sufficient for uniquely identifying the graph:*

(i) *For all* $j \in V$, $\sigma_j^2 = \sigma^2$ *for some* $\sigma^2 > 0$,

(i) $\sigma_j^2 \leq \sigma_k^2$ *where* j *comes before* k *in the ordering.*

(iii) $\min_{j \in V} \min_{k \in Pa(j)} \beta_{jk}^2 \geq \frac{\sigma_{\max}^2 - \sigma_{\min}^2}{\sigma_{\min}^2}$

(iv) $\min_{j \in V} \min_{k \in Pa(j)} \beta_{jk}^2 \geq \sigma_{\min}^2 \geq 1$.

We provide the proof in Supplementary. The conditions (i) and (ii) in Proposition 2.3 is implied by both Conditions (A) and (B) in Theorem 2.1. However, the Condition (iii) is only implied by Conditions (A), the Condition (iv) is only implied by Conditions (B) (see also Proposition 1 in **?**). Hence, Proposition 2.3 shows that if the equal variance assumption or monotone increasing error variance assumption holds, both forward selection and backward elimination approaches can recover the ordering. However, under the strictly monotone decreasing variances condition, the minimum value of edge weights should be sufficiently larger according to the error variances.

Now, we investigate the relationship between the conditions (A) and (B) in Theorem 2.1 using a simple 3-node chain graph. Consider a Gaussian SEM, $X_1 \to X_2 \to X_3$ such that

$$X_1 = \epsilon_1, \qquad X_2 = \beta_1 X_1 + \epsilon_2, \qquad X_3 = \beta_2 X_2 + \epsilon_3,$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ for all $j \in \{1, 2, 3\}$. Then, Assumption (A) is equivalent to the following three conditions:

$$(i)\ \ \sigma_2^2/\sigma_1^2 > (1 - \beta_1^2), \quad (ii)\ \ \sigma_3^2/\sigma_2^2 > (1 - \beta_2^2), \quad (iii)\ \ \frac{\sigma_3^2}{\sigma_1^2} > 1 - \beta_2^2\beta_1^2 - \frac{\beta_2^2\sigma_2^2}{\sigma_1^2}.$$

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

$$(i)\ \ \sigma_2^2/\sigma_1^2 > (1 - \beta_1^2), \quad (ii)\ \ \sigma_3^2/\sigma_2^2 > (1 - \beta_2^2), \quad (iii')\ \ \frac{\sigma_3^2}{\sigma_1^2} > 1 - \frac{\beta_1^2\sigma_3^2}{\sigma_2^2}.$$

Algorithm 1: Uncertainty Scoring Algorithm: Forward Selection

Input : n i.i.d. samples from a Gaussian Linear SEM, $X^{1:n}$

Output: Estimated causal graph, $\widehat{G} = (V, \widehat{E})$

Step (1): Ordering Estimation (Forward Selection);

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 conditional variance $\widehat{\sigma}^2_{j|S}$ by regressing X_j over X_S;

  end

  The m-th element of the ordering $\widehat{\pi}_m = \arg\min_j \hat{\sigma}^2_{j|S}$

end

Step (2): Parents Estimation;

for $m = \{2, \cdots, p\}$ **do**

  for $j = \{1, \cdots, m-1\}$ **do**

    Perform an independence test between $\widehat{\pi}_m$ and $\widehat{\pi}_j$;

    If dependent, include j into $\widehat{\text{Pa}}(\widehat{\pi}_m)$;

  end

end

Estimate the edge set $\widehat{E} := \cup_{m \in V} \cup_{k \in \widehat{\text{Pa}}(\widehat{\pi}_m)} (k, m)$;

Since the first two conditions are identical, we concentrate on the last conditions (iii) and (iii'').

Suppose that $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (2, 2, 1)$ and $\beta = (\beta_1, 1)$. Then, Assumption (B) is violated $1/2 \leq 1 - \beta_1^2/2$ if $\beta_1^2 \leq 1$ while Assumption (A) holds. In contrast, suppose that $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (2, 1.5, 1)$ and $\beta = (1, \beta_2)$. Then Assumption (A) is violated $1/2 \leq 1 - \beta_2^2\beta_1^2 - \frac{\beta_2^2\sigma_2^2}{\sigma_1^2} = 1 - \beta_2^2 - 3\beta_2^2/4$ if $\beta_2^2 \leq 2/7$ while Assumption (B) holds. Therefore, we cannot conclude that which assumption is strictly weaker.

3 Algorithm

In this section, we present our Variance Scoring algorithm for learning a Gaussian SEM (4). Algorithm 1 consists of two steps: (1) ordering estimation using the conditional variances; and (2) parent estimation using the (conditional) independence relations between variables. Our algorithm runs with any conditional variance estimation method and independence test.

Now, we present our choice of method for each step. Regarding the ordering estimation in Step (1), Algorithm 1 requires computation of conditional variances. Hence, we use a consistent estimator for the error variances using linear regression. More precisely, for $\text{Var}(X_j \mid X_S)$, we first regress X_j over X_S, and

Algorithm 2: Uncertainty Scoring Algorithm: Backward Elimination

Input : n i.i.d. samples from a Gaussian Linear SEM, $X^{1:n}$

Output: Estimated causal graph, $\widehat{G} = (V, \widehat{E})$

Step (1): Ordering Estimation (Backward Elimination);

Set $S = \{1, 2, \cdots, p\}$

for $m = \{p, p-1, \cdots, 1\}$ **do**

  for $j \in S$ **do**

    Estimate conditional variance $\widehat{\sigma}^2_{j|S}$ by regressing X_j over $X_{S \setminus j}$;

  end

  The m-th element of the ordering $\widehat{\pi}_m = \arg\min_j \hat{\sigma}^2_{j|S}$

  Update $S = S \setminus \pi_m$;

end

Step (2): Parents Estimation;

for $m = \{2, \cdots, p\}$ **do**

  for $j = \{1, \cdots, m-1\}$ **do**

    Perform an independence test between $\widehat{\pi}_m$ and $\widehat{\pi}_j$;

    If dependent, include j into $\widehat{\text{Pa}}(\widehat{\pi}_m)$;

  end

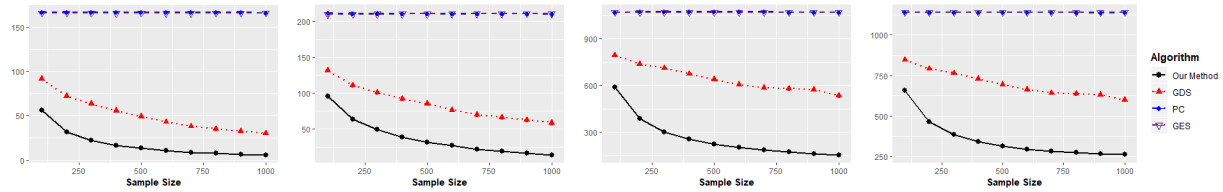
end

Estimate the edge set $\widehat{E} := \cup_{m \in V} \cup_{k \in \widehat{\text{Pa}}(\widehat{\pi}_m)} (k, m)$;

then, estimate $\text{Var}(X_j \mid X_S)$ using the residuals. Under the assumption in Theorem 2.1, the conditional variance of the correct element of the ordering π_j given $\pi_1, ..., \pi_{j-1}$ is strictly smaller than that of the other nodes in population. Hence, we can choose the correct element of the ordering with the smallest conditional variance. For the next element of the ordering π_{j+1}, we compute all conditional variances given $\pi_1, ..., \pi_j$. Therefore, the ordering is determined one node at a time by selecting the node with the minimum conditional variance and updating the condition set.

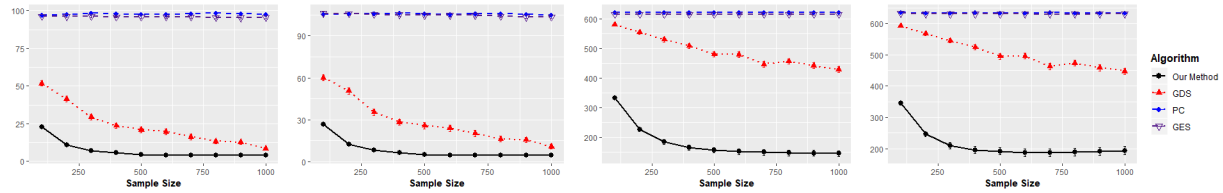
Estimating the set of parents of a node j in Step (2) boils down to selecting the parents among all elements before a node j in the ordering. Hence, given the estimated ordering from Step (1), Step (2) is reduced to a neighborhood selection problems using conditional dependence relations like the PC algorithm. However, unlike the PC algorithm which requires the faithfulness assumption, in our case, causal minimality is sufficient due to the ordering estimation in Step (1). As discussed, we do not assume causal minimality, but it is naturally mounted in our settings. We empirically verify that our algorithm does not need the faithfulness assumption in Section 4.3.

Compared to the greedy DAG search algorithm in Peters and Bühlmann (2014), another novelty of our algorithm is a polynomial-time complexity. More precisely, Peters and Bühlmann (2014) exploits the



(a) DAG: $p = 20$ (b) MEC: $p = 20$ (c) DAG: $p = 50$ (d) MEC: $p = 50$

Figure 2: Average Hamming distances between the estimated and true graphs, and the estimated and true MECs, when error variances are the same



(a) DAG: $p = 20$ (b) MEC: $p = 20$ (c) DAG: $p = 50$ (d) MEC: $p = 50$

Figure 3: Average Hamming distances between the estimated and true graphs, and the estimated and true MECs, when error variances are different

ℓ_0-penalized regression, and hence, computational cost grows super-exponentially as the number of nodes increases. In contrast, our method applies linear regression without any penalty terms, and conditional independence tests. Therefore, by decoupling the ordering estimation or parents search, we gain significant computational improvements. Similar ideas on reducing computational complexity by separating estimation of the ordering with the parents were applied in some existing algorithms (e.g., Bühlmann et al. 2014; Park and Raskutti 2018). We present the average run-time of both algorithms in Section 4.4.

4 Numerical Experiments

We provide simulation results to support our main theoretical results in Theorem 2.1 with various settings: (i) the same error variances, (ii) different error variances, and (iii) non-faithful distributions. We compared Algorithm 1 to state-of-the-art DAG learning PC, GES, and GDS algorithms in terms of the Hamming distance between the true and estimated graphs as in Peters and Bühlmann (2014). As we discussed, the PC and GES algorithms can learn only up to the MEC under the faithfulness assumption. Hence, we also report the Hamming distance between the true and estimated MECs.

Step (2) of Algorithm 1 and the PC algorithm were implemented using a Fisher's exact independence test. In addition, ignoring multiple testing issues, we always set the significance level of statistical tests to $\alpha = 1\%$. The GES algorithm exploits the BIC-regularized maximum likelihood of Gaussian SEMs. Lastly,

for GDS, we set the initial graph to the empty graph. Since the GDS algorithm uses a greedy search, and its accuracy relies on the initial graph, we acknowledge that GDS can be better with prior information of an initial graph.

4.1 Random Gaussian SEMs with Homogeneous Error Variances

We conducted simulations using 100 realizations of p-node Gaussian SEMs (4) with a randomly generated underlying DAG structure. The set of parameters $\beta_{jk} \in \mathbb{R}$ in Equation (4) was generated uniformly at random in the range $\beta_{jk} \in [-2, 2]$, and was then set to 0 if $\beta_{jk} \in (-0.25 \cup 0.25)$. Hence, the graphs we considered may not be sparse. Lastly, all noise variances were set to 1.

In Fig. 2, we compare our algorithm to state-of-the art PC, GES, and GDS algorithms by varying sample size $n \in \{100, 200, ..., 1000\}$ and node size $p \in \{20, 50\}$. As expected, Fig. 2 shows that our algorithm and GDS consistently recover the true graph, and hence, we empirically verify that Gaussian SEMs with identical errors are identifiable. In addition, our method outperforms the 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 2.1. Lastly, the PC and GES algorithms seem to fail to recover both directed graphs and the MECs. It is worth noting that the PC and GES algorithms are not consistent, and often fail to recover the MEC if a true graph is not sparse due to the very strong faithfulness assumption in finite samples Uhler et al. (2013).

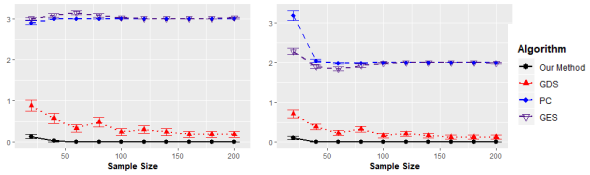
4.2 Random Gaussian SEMs with Heterogeneous Error Variances

We generated 100 sets of samples with the same procedure specified in Section 4.1, except that randomly chosen error variances, $\sigma_j^2 \in [1, 3]$, and the range of parameters, $\beta_{jk} \in [-2, 2]$ and was set to 0 if $\beta_{jk} \in (-1, 1)$. We note that this range of parameters, β_{jk}, forces the graphs to be sparser and ensures that our identifiability assumption in Theorem 2.1 is satisfied with any values of error variances.

In Fig. 3, we evaluated Algorithm 1 and the comparison methods by varying sample size $n \in \{100, 200, ..., 1000\}$ and node size $p \in \{20, 50\}$. Fig. 3 shows that our algorithm consistently recovers the true graph, and therefore, confirms our theoretical findings that Gaussian SEMs are identifiable, even with different error variances. Fig. 3 also shows that the GDS algorithm recovers graphs more accurately as a sample size increases. This robustness to non-identical errors is not a surprising result, according to Section 5.3 in Peters and Bühlmann (2014), although they do not provide legitimate reasons. Lastly, the PC and GES algorithms still show poor performances when learning MECs in our settings.

4.3 Non-faithful Gaussian SEMs with Heterogeneous Error Variances

In Section **??**, we proved that Gaussian SEMs are identifiable even when the distributions are non-faithful. Hence, in this section, we empirically verify this phenomenon. We generated 100 sets of samples from the



(a) DAG: $p = 3$ (b) MEC: $p = 3$

Figure 4: Average Hamming distances between the estimated and true graphs, and the estimated and true MECs, when the distribution is non-faithful and error variances are different

following non-faithful directed graphical models:

$$X_1 = \epsilon_1, \quad X_1 = X_1 + \epsilon_2, \quad X_3 = X_1 + X_2 + \epsilon_3,$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ has $\sigma_1^2 = 2.25$ and $\sigma_2^2 = \sigma_3^2 = 1.5$. We note that it is not a very favorable setting for our algorithm, since $\sigma_1^2 > \sigma_j^2$ for all $j \in \{2, 3\}$.

Fig. 4 compares the DAG learning algorithm as a function of sample size $n \in \{20, 40, ..., 200\}$. As we can see in Fig. 4, it confirms that our algorithm and GDS do not require the faithfulness assumption to recover the underlying graphs of Gaussian SEMs. Fig. 4 also shows that our algorithm performs better than the comparison algorithms on average.

4.4 Computational Complexity

One of the important issues in learning DAG models is computational complexity due to the super-exponentially growing size of the space of DAGs in the number of nodes (Harary, 1973). Hence, it is in general NP-hard to search DAG space (Chickering et al., 1994; Chickering, 1996), and many existing algorithms, such as PC, GES, MMHC, and GDS, are inevitably heuristic, which may not guarantee recovering the true graph. Hence, we now investigate the run-time of our algorithm using the random Gaussian DAG models with identical errors discussed in Section 4.1.

Table 1 compares the run-time of Algorithm 1 to GDS for learning Gaussian SEMs by varying sample size $n \in \{100, 200, ..., 1000\}$ and node size $p \in \{20, 50, 80\}$. Table 1 shows that our algorithm is computationally feasible, even in large-scale graphs learning. In particular, our algorithm is almost 400 times faster than GDS when $p = 20$. As the node size gets bigger, our algorithm is even faster than GDS, and is approximately 800 times faster when $p = 50$. We do not apply GDS when $p = 80$ due to a very long run-time that takes more than a day if implemented. Again, we emphasize that our mild identifiability assumption enables our algorithm to be a large-scale graph learning algorithm.

Table 1: Comparison of our algorithm (denoted OUR) to the GDS algorithm in terms of average run-time (in seconds) with respect to node size p and sample size n

	$p = 20$		$p = 50$		$p = 80$
n	OUR	GDS	OUR	GDS	OUR
100	0.63	233.21	5.07	2617.53	17.01
200	0.68	268.20	5.74	3504.60	20.30
300	0.72	302.58	6.71	4284.35	24.76
400	0.75	325.94	7.24	4832.54	27.07
500	0.83	355.60	8.39	5734.61	32.18
600	0.84	370.94	8.61	6109.63	33.61
700	0.90	396.83	9.97	7047.25	39.48
800	0.94	412.48	10.16	7517.26	40.36
900	1.01	436.60	11.57	8234.65	46.70
1000	1.01	448.28	11.62	8695.12	46.87

5 Discussion

We proved the identifiability of Gaussian SEMs with both identical and non-identical errors only from joint distribution using the uncertainty level of the conditional variables. Our approach requires commonly assumed causal sufficiency, the non-zero edge weights assumption, and the new identifiability assumption in Theorem 2.1. We assume neither causal minimality nor faithfulness that can be very restrictive. Based on our identifiability assumption, we propose a statistically consistent and computationally feasible algorithm. Our algorithm can be implemented with any combination of conditional variance estimation methods and independence tests. In addition, it can be applied to high-dimensional data using ℓ_1-regularized regression if a graph is sparse. Moreover, our theoretical findings can be combined with an existing MEC learning algorithm for recovering the causal graph, because our method can estimate the ordering independent of directed edges or skeleton.

References

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

David M Chickering, Dan Geiger, David Heckerman, et al. Learning bayesian networks is np-hard. Technical report, Citeseer, 1994.

David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

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

Kenji Doya. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007.

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.

Frank Harary. New directions in the theory of graphs. Technical report, MICHIGAN UNIV ANN ARBOR DEPT OF MATHEMATICS, 1973.

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.

Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pages 343–359. IEEE, 1991.

Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

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.

Gunwoong Park. Learning generalized hypergeometric distribution (ghd) dag models. *arXiv preprint arXiv:1805.02848*, 2018.

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

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.

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.

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

6 Appendix

6.1 Proof for Theorem 2.1

Proof. Without loss of generality, we assume that the true ordering $\pi = (\pi_1, ..., \pi_p)$ is unique. For simplicity, we define $X_{1:j} = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_j})$ and $X_{1:0} = \emptyset$. We restate the identifiability assumption of Gaussian SEMs.

Assumption 6.1 (Identifiability)**.** For any node $m \in V$, let $j = \pi_m$ and $k \in V \setminus \text{Nd}(j)$. The conditional variance of X_j given its parents is smaller than the conditional variance of X_k given the variables before j in ordering π:

$$\sigma_j^2 < \sigma_k^2 + \mathbb{E}(\text{Var}(\mathbb{E}(X_k \mid X_{\text{Pa}(k)}) \mid X_{1:(j-1)})).$$

Now, we prove identifiability of Gaussian SEMs using mathematical induction.

Step (1) By Assumption 6.1, for any node $k \in V \setminus \{\pi_1\}$, we have

$$\text{Var}(X_{\pi_1}) = \sigma_{\pi_1}^2 < \sigma_k^2 + \text{Var}(\mathbb{E}(X_k \mid X_{\text{Pa}(k)})) = \text{Var}(X_k).$$

Therefore, π_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 estimated.

Step (m) Now, we consider the m^{th} element of the causal ordering and its parents. By Assumption 6.1, for $k \in \{\pi_{m+1}, \cdots, \pi_p\}$,

$$\mathbb{E}(\text{Var}(X^2_{\pi_m} \mid X_{1:(m-1)})) = \sigma^2_{\pi_m} < \sigma^2_k + \mathbb{E}(\text{Var}(\mathbb{E}(X_k \mid X_{\text{Pa}(k)}) \mid X_{1:(m-1)})) = \mathbb{E}(\text{Var}(X_k \mid X_{1:(m-1)})).$$

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

In terms of the parents search, it is clear that conditional independence relations naturally encoded by the factorization (1) and imply causal minimality (see details in Pearl 2014; Peters and Bühlmann 2014). In our settings, causal minimality states that for any node $j \in V$ and one of its parents $k \in \text{Pa}(j)$,

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

Therefore, we can choose the correct parents of π_m. By mathematical induction, this completes the proof. □

6.2 Identifiability for Three-node Chain Graph

Consider a Gaussian SEM, $X_1 \to X_2 \to X_3$, where $X_1 = \epsilon_1$, $X_2 = \beta_1 X_1 + \epsilon_1$, and $X_3 = \beta_2 X_2 + \epsilon_3$ with $\epsilon_j \sim N(0, \sigma^2_j)$ for all $j \in \{1, 2, 3\}$. Then the first element of the ordering can be determined by comparing the variances of nodes:

$$\begin{aligned}
\text{Var}(X_2) &= \mathbb{E}(\text{Var}(X_2 \mid X_1)) + \text{Var}(\mathbb{E}(X_2 \mid X_1)) = \sigma^2_2 + \beta^2_1 \sigma^2_1 > \sigma^2_1 = \text{Var}(X_1) \\
\text{Var}(X_3) &= \mathbb{E}(\text{Var}(X_3 \mid X_2)) + \text{Var}(\mathbb{E}(X_3 \mid X_2)) = \sigma^2_3 + \beta^2_2 \sigma^2_2 + \beta^2_2 \beta^2_1 \sigma^2_1 > \sigma^2_1 = \text{Var}(X_1).
\end{aligned}$$

as long as $\sigma^2_2/\sigma^2_1 > (1 - \beta^2_1)$ and $\sigma^2_3/\sigma^2_1 > (1 - \beta^2_2)$.

The second element of the ordering can also be recovered by comparing the expectation of the conditional variance of the remaining variables given the estimated first element of the ordering:

$$\begin{aligned}
\mathbb{E}(\text{Var}(X_3 \mid X_1)) &= \mathbb{E}(\mathbb{E}(\text{Var}(X_3 \mid X_2) \mid X_1)) + \mathbb{E}(\text{Var}(\mathbb{E}(X_3 \mid X_2) \mid X_1)) \\
&= \sigma^2_3 + \beta^2_2 \sigma^2_2 > \sigma^2_2 = \mathbb{E}(\text{Var}(X_2 \mid X_1)).
\end{aligned}$$

as long as $\sigma^2_3/\sigma^2_2 > (1 - \beta^2_2)$.

Under the minimality and the Markov condition, we also have the following (conditional) dependence relations: $X_1 \not\perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_3$, $X_2 \not\perp\!\!\!\perp X_3$ Therefore, the true graph can be recovered.