Scalable Intervention Target Estimation in Linear Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper considers the problem of estimating the unknown intervention targets in a causal directed acyclic graph from observational and interventional data generated under soft interventions and linear structural equation models (SEMs). Current causal structure learning methods either work with known intervention targets or use hypothesis testing to discover the unknown intervention targets even for linear SEMs, limiting their scalability and sample efficiency. This paper proposes a scalable and efficient algorithm that consistently identifies all intervention targets. The pivotal idea is to estimate the intervention sites from the difference in precision matrices, across observational and interventional datasets, by repeatedly applying them on different subsets of variables. This approach facilitates establishing sample complexity guarantees as well as scalability to larger graphs. The proposed algorithm can be used to also update a given observational Markov equivalence class into the interventional Markov equivalence class. Consistency, Markov equivalency, and sample complexity are established analytically. Finally, simulation results on both real and synthetic data demonstrate the gains of the proposed approach for scalable causal structure recovery.

1 Introduction

Directed acyclic graphs (DAG) are the most common models for encoding the cause-effect relationships among random variables. Extensive research has been dedicated to learning the structure of DAGs from their associated observed data. Structure learning from the observational data relies on uncovering conditional independence (CI) among the random variables. Since structurally distinct DAGs can encode the same set of CI relations, a DAG is identifiable only up to its Markov equivalence class (MEC) from the observational data. Subsequently, *interventional* data can be used to refine the MEC obtained from the observational data and learn specific causal effects.

This paper is motivated by addressing two significant independent challenges in causal discovery. First, most of the existing approaches for learning with interventional datasets require the *intervention target set* to be known, which can be a strong assumption. For instance, gene-editing technologies are known to perform cleavage at off-target genome sites [1]. Therefore, identifying the *intervened* nodes alone is a critical problem in structure learning, and despite its significance, it remains uninvestigated. Secondly, besides learning the structures of single DAGs, there exist application domains in which the goal is instead learning the structural changes between two related networks and their associated DAGs. For instance, structural differences between the gene regulatory networks of different subtypes of cancers may help to identify the roles of specific genes [2].

In practice, fixing the target variable at a specific value or removing its causal dependencies is often difficult, but the distribution of the target node can instead be modified [3]. Therefore, we consider a *soft intervention* setting, in which we assume the variances of the noise terms for targets are changed. Furthermore, we assume that the topological ordering remains the same after the intervention.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Under this soft intervention model, we propose an algorithm for estimating the intervention targets given the data from two linear SEMs, i.e., observational and interventional. We build on estimating the sparse difference between precision matrices to avoid the extensive conditional independence tests. This leads to significant improvement over the computational complexity of alternative methods that can be suitable modified for intervention recovery tasks. This facilitates scaling up to high-dimensional settings. Furthermore, we show that this algorithm can be used in conjunction with an observational DAG learning procedure to refine the MEC to *interventional*-MEC ($\mathcal{I}$-MEC). Besides being consistent in the population setting, we provide the finite-sample guarantees for linear SEM with Gaussian noise when the soft interventional changes between the two models are sparse. Our main contributions are as follows:

- We propose an algorithm that identifies intervention targets under the intervention-faithfulness assumption. We show that our algorithm identifies $\mathcal{I}$-MEC given the observational MEC.
- We provide the sample complexity results of our algorithm under linear SEM with Gaussian noise.
- We perform experiments on both real biological and synthetic datasets to illustrate the ability to work in high-dimensional settings and the gains compared to the relevant methods.

2 Related work

Among the broad range of approaches to intervention recovery, there exist two methods that are closely related to the scope of this paper: (i) estimating the difference between two DAGs and (ii) learning from a combination of observational and interventional datasets.

Direct estimation of differences: Direct estimation of differences in linear SEMs has been studied recently. The study in [4] proposes a PC-style algorithm for learning changes in the edge weights by testing invariances of regression coefficients and noise variances. Even though the differences can be sparse, individual models can be dense, and estimating these variables through regression can be inaccurate. Furthermore, the number of hypothesis tests is exponential in the number of nodes that are affected by the changes, and this can be prohibitive even in sparse changes on hub nodes. Estimating the difference of two precision matrices is a relatively easier task and has received attention lately [5–8]. When such a difference is sparse, these methods provide finite-sample guarantees in the high-dimensional regime. An existing study closest to the scope of our work is [9], which proposes re-estimating precision difference to progressively eliminate nodes and estimate the difference DAG. This approach critically hinges on the assumption that the noise variance is invariant, rendering its applicability to intervention settings. In contrast, our algorithm builds on the changes in noise variances and estimates the intervention targets efficiently. We demonstrate the effect of this difference in Appendix B.1.

Learning interventional-MEC: There is a growing number of studies on causal structure learning from both observational and interventional data. Score-based greedy interventional equivalence search (GIES) [10] and hybrid interventional greedy sparsest permutation (IGSP) [11] algorithms are proposed when there are no latent confounders. Both of these studies assume that the intervention targets are known. However, knowing the intervention targets can be a strong assumption since even for controlled interventional experiments, off-target effects are common. For instance, noise variances of the off-target variables can change, resulting in the intervened nodes being unknown.

For the causal structure learning without knowledge of the intervention targets, an early study includes the dynamic programming approach of [12]. This approach has limited scalability due to its time complexity being exponential in the model size. Structural discovery from interventions proposed in [13] is a neural network-based method that can learn from interventional data without target knowledge. Nevertheless, it requires discrete data and can have at most one intervened node in a given setting, limiting its applicability. The differential causal discovery from interventional data (DCDI) algorithm proposed in [14] extends the approach of [13] without making as strong assumptions, e.g., learning a distribution over all potential interventional families via continuous optimization. While this method is shown to converge to the real intervention targets, its runtime becomes prohibiting in models with as few as 100 nodes. The Unknown-target IGSP (UT-IGSP) algorithm proposed in [15] learns the intervention targets simultaneously while learning the causal structure. Even though interventions refine the search space, a greedy search of the sparsest permutation is still too slow in the high-dimensional regime, especially when using non-Gaussian CI tests.

3 Problem definition

Let $\mathcal{G} \triangleq ([p], E)$ be a DAG with the node set $[p] := \{1, \dots, p\}$ and the edge set E. We denote the directed edge from $i \in [p]$ to $j \in [p]$ by $i \to j$. We associate a random variable X_i with $i \in [p]$, and define the random vector $X \triangleq (X_1, \dots, X_p)^\top$. We consider the linear SEM according to which

$$X = B^\top X + \epsilon \,, \tag{1}$$

where $B \in \mathbb{R}^{p \times p}$ is the autoregressive matrix in which $B_{ij} \neq 0$ if and only if $i \to j$ in $\mathcal{G}$. The random $\epsilon \in \mathbb{R}^{p \times 1}$ is zero-mean and with covariance matrix $\Omega := \mathsf{diag}(\sigma_1^2, \dots, \sigma_p^2)$. We denote the covariance matrix of X by Σ and its inverse (the precision matrix) by Θ, which satisfies $\Theta = (I - B)\Omega^{-1}(I - B)^\top$. For entries of Θ, we have

$$\Theta_{ij} = -\frac{B_{ij}}{\sigma_j^2} - \frac{B_{ji}}{\sigma_i^2} + \sum_{k \in \mathsf{ch}(i) \cap \mathsf{ch}(j)} \frac{B_{ik} B_{jk}}{\sigma_k^2} \,, \quad \forall i \neq j \,, \tag{2}$$

$$\Theta_{ii} = \sigma_i^{-2} + \sum_{j \in \mathsf{ch}(i)} \sigma_j^{-2} B_{ij}^2 \,, \quad \forall i \in [p] \,, \tag{3}$$

where $\mathsf{ch}(i), \mathsf{pa}(i), \mathsf{de}(i), \mathsf{an}(i)$ denote children, parents, descendants, and ancestors set of node i in DAG $\mathcal{G}$, respectively.

From the observational data, a DAG can be learned only up to its MEC [16]. Interventions are used to increase the identifiability of a DAG by removing all causes of the intervention target (*perfect intervention*) or modifying those relationships without removing them completely (*imperfect intervention*). We consider the following *soft intervention* setting, which does not remove causal effects (from direct parents) on intervention target nodes, and hence, is more practical.

Soft intervention model. In this model, interventions correspond to disturbing the target nodes $i \in \mathcal{I}$ by changing the variances of their noise variables whereas not necessarily changing the weights of the causes, i.e., weights $B_{\mathsf{pa}(i),i} \triangleq \{B_{j,i} : j \in \mathsf{pa}(i)\}$ can vary freely. This intervention procedure on the initial DAG results in a second DAG with new parameters.

Let $\mathcal{G}^{(1)}$ represent a linear SEM with $(B^{(1)}, \epsilon^{(1)})$ and $\mathcal{G}^{(2)}$ be the linear SEM after the intervention, with $(B^{(2)}, \epsilon^{(2)})$ parameters. The intervention target set that relates these two DAGs is

$$\mathcal{I} \triangleq \{i : \sigma_i^{(1)} \neq \sigma_i^{(2)}\} \,. \tag{4}$$

Accordingly, denote the covariance and precision matrices of these models with $\Sigma^{(1)}, \Sigma^{(2)}, \Theta^{(1)}, \Theta^{(2)}$, and differences between the two models by $\Delta_B \triangleq B^{(1)} - B^{(2)}$ and $\Delta_\Theta \triangleq \Theta^{(1)} - \Theta^{(2)}$. For a subset of nodes $S \subseteq [p]$, denote the precision matrix of random vector $X_S \triangleq \{X_i : i \in S\}$ by Θ_S. We also define the set of changed nodes as $S_\Delta \triangleq \{k : (\Delta_\Theta)_{k,k} \neq 0\}$ and denote its size by $p_\Delta \triangleq |S_\Delta|$. According to (3), this set consists of all the intervened nodes and their parents.

In this paper, we estimate intervention targets $\mathcal{I}$ given the data from two SEMs under soft intervention model. Furthermore, we estimate the non-intervened parents of the targets, which updates the given observational MEC into the $\mathcal{I}$-MEC. Mathematically, we define $\psi : \Sigma^{(1)} \times \Sigma^{(2)} \to \hat{\mathcal{I}}, \mathsf{pa}(\hat{\mathcal{I}})$ as the estimator that maps the covariance matrices of observational and interventional data to an intervention target set estimate and their parents. We aim to maximize the following probabilities that ψ exactly recovers $\mathcal{I}$ and their parents

$$\mathsf{P} \triangleq \mathbb{P}(\mathcal{I} = \hat{\mathcal{I}}) \,, \quad \mathsf{Q} \triangleq \mathbb{P}(\mathsf{pa}(\mathcal{I}) = \mathsf{pa}(\hat{\mathcal{I}})) \,. \tag{5}$$

We will show that the algorithm proposed in Section 4 exactly recovers $\mathcal{I}$ and their non-intervened parents, and hence, given the observational MEC, we obtain the $\mathcal{I}$-MEC.

4 Algorithm and main results

In this section, we provide our proposed algorithm and present the attendant performance guarantees. Our algorithm builds on repeatedly estimating the difference of precision matrices to find the intervention targets $\mathcal{I}$, or equivalently, its complement $\mathcal{I}^{\mathrm{C}}$. The algorithm consists of three key steps.

In **Step 1**, instead of directly estimating $\mathcal{I}$, we aim to discard the nodes that are deemed strongly not to belong to $\mathcal{I}$. For this purpose, we start by identifying the non-intervened nodes that do not have intervened children. These nodes are not of our interest and discarded from further consideration and we continue with only the nodes contained in S_Δ for further scrutiny. A naive approach to identifying the rest of the non-intervened nodes will be computing Δ_Θ exhaustively for all the 2^{p_Δ} subsets of S_Δ, which for large p_Δ can be computationally prohibitive. Alternatively, we partition S_Δ to two sets: the set of *non-intervened source* nodes and its complement, and feed these two partitions for further processing to Step 2. Since distribution of a node relies only on its ancestors, reaching a topological ordering is critical to reduce complexity. In **Step 2**, we find the descendants of the source nodes and decompose S_Δ into equivalence classes according to their ancestors in non-intervened source nodes. Decomposing S_Δ allows us to order the equivalence classes according to a topological ordering. We process these classes in **Step 3** individually, and show that we only need to compute Δ_Θ for all the subsets of each class considered in Step 3. This results in a significant reduction in the computational complexity compared to the exhaustive search approach. Finally, we identify non-intervened parents of $\mathcal{I}$ from the earlier results. Since precision difference estimation function appears in our algorithm repeatedly, we describe it next and continue with the details of the three steps.

Precision difference estimation (PDE). When the differences between two SEMs are sparse, estimating the difference of their precision matrices can be formulated as a Lasso-type problem and solved efficiently. Since it will be used repeatedly, it is important for this function to recover the support of Δ_Θ and have feasible computational complexity. The study in [7] solves the following convex problem through the alternating direction method of multipliers (ADMM) to estimate $\Delta = \Theta^{(1)} - \Theta^{(2)}$:

$$\hat{\Delta} = \underset{\Delta}{\operatorname{argmin}} \frac{1}{2} Tr(\Delta^\top \hat{\Sigma}^{(1)} \Delta \hat{\Sigma}^{(2)}) - Tr(\Delta(\hat{\Sigma}^{(1)} - \hat{\Sigma}^{(2)})) + \lambda \|\Delta\|_1 \ , \tag{6}$$

where λ is a tuning parameter. The computational complexity of this algorithm grows according to $O(p^3)$, which overcomes the limitation of the algorithm of [5] used in [9], which has complexity $O(p^4)$. Therefore, we use this ADMM-based method as our PDE function.

Step 1: Finding non-intervened source nodes. By levering the PDE procedure discussed, we first estimate Δ_Θ over all $[p]$ nodes and obtain S_Δ. Representation of $\Theta_{i,i}$ in (3) implies that a diagonal entry of Δ_Θ is non-zero if and only if the corresponding node is in $\mathcal{I}$ or it has a child in $\mathcal{I}$. Then, $[p] \setminus S_\Delta$ set contains only non-intervened nodes and can be discarded from consideration. Furthermore, some non-intervened nodes in S_Δ, namely those that do not have an intervened ancestor, do not observe a change in their statistics. Therefore, it is possible to identify them from covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$, and subsequently, they can serve as the starting points for distinguishing the rest of the $\mathcal{I}^{\mathrm{C}}$ in S_Δ. Let us define them as the *non-intervened source nodes*

$$J_0 \triangleq \{j : j \in S_\Delta,\ j \notin \mathcal{I},\ \mathsf{an}_{\mathcal{I}}(j) = \emptyset\} \ . \tag{7}$$

The outputs of this step, S_Δ and J_0, are fed into the next steps of the algorithm.

Step 2: Forming equivalence classes from J_0. We will show that for any non-intervened node $j \in S_\Delta$, there exists a minimal subset of S_Δ, which makes the corresponding diagonal entry of the precision matrix invariant, and it does not contain any descendant of j. Therefore, the non-intervened nodes that have the same ancestors are affected by the interventions similarly, and finding their ancestors is critical. We show that determining whether node $k \in J_0$ is an ancestor of node $j \in S_\Delta \setminus J_0$ is possible by applying PDE on $\{j, k\}$. Accordingly, we define the *source ancestor set* J_0^k for each node $k \in S_\Delta \setminus J_0$, which consists of the ancestors of k in set J_0,

$$J_0^k \triangleq \{j : j \in J_0 \cap \mathsf{an}(k)\}, \quad \forall k \in S_\Delta \setminus J_0 \ . \tag{8}$$

Next, we decompose the set $S_\Delta \setminus J_0$ into *equivalence classes* where all the nodes in a class have the same source ancestor set, i.e., they share the same ancestors in J_0. We denote these equivalence classes by $\mathcal{A}_1, \ldots, \mathcal{A}_L$, and the source ancestor set corresponding to a class $\mathcal{A}_\ell$ by $J_0^{\mathcal{A}_\ell}$ for $\ell \in [L]$. These classes are ordered according to topological ordering such that for $1 \leq \ell < \ell' \leq L$, $J_0^{\mathcal{A}_{\ell'}} \not\subset J_0^{\mathcal{A}_\ell}$. In other words, the class corresponding to the superset of any $J_0^{\mathcal{A}_\ell}$ should appear later than $\mathcal{A}_\ell$ in the sequence $\mathcal{A}_1, \ldots, \mathcal{A}_L$. Source ancestor sets and equivalence classes are fed into the next step.

Step 3: Processing equivalence classes. Given Steps 1 and 2, we can describe our exact search space of subsets for Δ_Θ estimates to declare whether a node is intervened. We process equivalence classes $\mathcal{A}_1, \ldots, \mathcal{A}_L$ individually, i.e., at stage ℓ, we consider the nodes in $\mathcal{A}_\ell$. Let us define

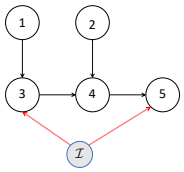
$\mathcal{M}_\ell \triangleq \bigcup_{1\le b<\ell} \mathcal{A}_b$. It suffices to estimate $\Delta_{\Theta_{\mathcal{M}_\ell \cup A}}$ **only for each** $A \subseteq \mathcal{A}_\ell$ to determine the intervention status of any node in $\mathcal{A}_\ell$ class. This key observation reduces the number of PDE steps needed. Specifically, for any non-intervened $j \in \mathcal{A}_l$ there exists a subset $A \subseteq \mathcal{A}_\ell$ such that the corresponding diagonal entry of $\Delta_{\Theta_{\mathcal{M}_\ell \cup A}}$ will be zero, and there does not exist any such set for intervened nodes in $\mathcal{A}_\ell$. Formally, the *process equivalence class* returns

$$\mathcal{I}_\ell = \{i : i \in \mathcal{A}_\ell \cap \mathcal{I}\}\,, \quad \text{and} \quad J_\ell = \{j : j \in \mathcal{A}_\ell \cap \mathcal{I}^{\mathrm{C}}\}\,. \tag{9}$$

Finally, we identify the non-intervened parents of intervened nodes without any new Δ_Θ estimates.

Computational complexity. Algorithm 1 repeatedly uses PDE in each step. The number of required instances of PDE is $(p_\Delta + 1)$ in Step 1, $O(p_\Delta^2)$ in Step 2, and $\sum_{\ell\in[L]} 2^{|\mathcal{A}_\ell|}$ in Step 3. Hence, it grows exponentially with $\max_{\ell\in[L]} |\mathcal{A}_\ell|$, which can be p_Δ in the worst case in extreme examples. Nevertheless, in almost all practical scenarios it is usually much smaller. To provide some insights, we provide the next example and more discussions in Appendix B.4.

Example 1. Consider a DAG with nodes $\{1, 2, 3, 4, 5\}$ and edge set $\{1 \to 3, 3 \to 4, 2 \to 4, 2 \to 5, 4 \to 5\}$ and let $\mathcal{I} = \{3, 5\}$. Then we have $S_\Delta = \{1, 2, 3, 4, 5\}$, $J_0 = \{1, 2\}$, $J_0^3 = \{1\}$, $J_0^4 = \{1, 2\}$, $J_0^5 = \{1, 2\}$, and accordingly, $\mathcal{A}_1 = \{3\}$, $\mathcal{A}_2 = \{4, 5\}$, $J_0^{\mathcal{A}_1} = \{1\}$, $J_0^{\mathcal{A}_2} = \{1, 2\}$. Note that largest $\mathcal{A}$ class has 2 nodes whereas S_Δ has 5 nodes.



Algorithm 1 Causal Intervention Target Estimator (CITE)

1: **Input:** $\Sigma^{(1)}$ and $\Sigma^{(2)}$
2: **Output.** $\mathcal{I}$ intervention set
3: Estimate $\Delta_\Theta \leftarrow$ *precision difference*$(\Sigma^{(1)}, \Sigma^{(2)})$
4: Form $S_\Delta := \{k : k \in [p],\ [\Delta_\Theta]_{k,k} \neq 0\}$
5: Form J_0 using (7) and J_0^k for each $k \in S_\Delta \setminus J_0$ using (8)
6: Form equivalence classes $\mathcal{A}_1, \ldots, \mathcal{A}_L$.
7: **for** $\ell \in [L]$ **do**
8: Take $\mathcal{A}_\ell$ set and the corresponding $J_0^{\mathcal{A}_\ell}$ set
9: $\mathcal{B}_\ell \leftarrow \{b : J_0^{\mathcal{A}_b} \subset J_0^{\mathcal{A}_\ell},\ \ 1 \le b < \ell\}$
10: $\mathcal{M}_\ell = J_0^{\mathcal{A}_\ell} \cup \bigcup_{b\in\mathcal{B}_\ell} \mathcal{A}_b$
11: $J_\ell, \mathcal{I}_\ell \leftarrow$ *process equivalence class*$(\mathcal{M}_\ell, \mathcal{A}_\ell, \Sigma^{(1)}, \Sigma^{(2)})$
12: **end for**
13: $\mathcal{I} = \bigcup_{\ell\in[L]} \mathcal{I}_\ell$
14: $\mathsf{pa}(\mathcal{I}) \leftarrow$ *parent finder*$(\mathcal{I}, \mathcal{J}, \mathcal{M}_1, \mathcal{A}_1, \ldots, \mathcal{M}_L, \mathcal{A}_L, \Sigma^{(1)}, \Sigma^{(2)})$
15: **Return** $\mathcal{I}$ and $\mathsf{pa}(\mathcal{I})$

Restricted SEM For a linear SEM $\mathcal{G}$ with (B, ϵ), we denote the restricted SEM (see Lemma 1 for details) that characterizes the relationship among the random variables X_S for a set S by (B_S, ϵ_S). As defined before, the corresponding precision matrix is denoted by Θ_S. The entries of B_S and noise variances σ_S are characterized by the original values of B, σ and Θ.

Remark 1 *We remark that invariance of the distributions for the noise term of a node and the value of the node are equivalent only for the non-intervened nodes that do not have an intervened ancestor. Therefore, only such nodes can be detected from the full linear SEM. The noise term of a non-intervened node maintains its invariance in a restricted SEM in which we remove all of its descendants and keep its intervened ancestors and their parents. On the other hand, noise term of an intervened node is always variant for any choice of restricted SEM.*

In the subsequent subsections, we will provide different analytical guarantees of Algorithm 1. Specifically, we will comment on the consistency of $\mathcal{I}$ recovery, refinement of MEC to $\mathcal{I}$-MEC, and sample complexity.

4.1 Consistency of $\mathcal{I}$ recovery

We provide the consistency results for Algorithm 1 at estimating $\mathcal{I}$ in this subsection. First, we need the following assumption to ensure that interventions are successful.

Algorithm 2 Functions for the main algorithm

Precision Difference Estimation (PDE) (Σ_1, Σ_2)

Using ADMM based algorithm of [7] method estimate $\Delta_\Theta = {\Sigma_1}^{-1} - {\Sigma_2}^{-1}$
Symmetrize Δ_Θ: set $\Delta_\Theta = (\Delta_\Theta + \Delta_\Theta^\top)/2$
Threshold Δ_Θ: set $[\Delta_\Theta]_{i,j} = 0$ if $|(\Delta_\Theta)_{ij}| < \varepsilon$.
Return Δ_Θ

Process Equivalence Class $(\mathcal{M}, \mathcal{A}, \Sigma_1,\ \Sigma_2)$

For each subset $A \subseteq \mathcal{A}$, estimate $\Delta_{\Theta_{\mathcal{M}\cup A}} \leftarrow$ *precision difference*$((\Sigma_1)_{A,A}, (\Sigma_2)_{A,A})$
for $k \in \mathcal{A}$ **do**
  if $\exists A \subseteq \mathcal{A}$, where $k \in A$, and $[\Delta_{\Theta_{\mathcal{M}\cup A}}]_{k,k} = 0$ **then**
    $J \leftarrow J \cup k$
  else
    $\mathcal{I} \leftarrow \mathcal{I} \cup k$
  end if
end for
Return $J, \mathcal{I}$

Parent Finder $(\mathcal{I}, \mathcal{J}, \mathcal{M}_1, \mathcal{A}_1, \ldots, \mathcal{M}_L, \mathcal{A}_L, \Sigma_1, \Sigma_2)$

for $i \in \mathcal{I}$ **do**
  $c_i \leftarrow c_i \triangleq \ell : i \in \mathcal{A}_\ell$
  for $j \in \mathcal{M}_i \cup J_{c_i}$ **do**
    if $\nexists A \subseteq \mathcal{A}_{c_i}$ such that $[\Delta_{\Theta_{\mathcal{M}_i\cup A}}]_{j,i} = 0$ **then**
      Add j to $\mathsf{pa}(i)$
    end if
  end for
end for
Return $\mathsf{pa}(i)$ for $i \in \mathcal{I}$

Assumption 1 **($\mathcal{I}$-faithfulness)** *For any choice of $i, j \in S \subseteq [p]$, we have the following properties:*

1. *If $\sigma_i^{(1)} \neq \sigma_i^{(2)}$, then $\sigma_{S,i}^{(1)} \neq \sigma_{S,i}^{(2)}$.*
2. *If $\sigma_{S,i}^{(1)} \neq \sigma_{S,i}^{(2)}$, then $[\Theta_S^{(1)}]_{i,i} \neq [\Theta_S^{(2)}]_{i,i}$.*
3. *If $[B_S]_{i,j} \neq 0$ in either model, then $[\Theta_S^{(1)}]_{i,j} \neq [\Theta_S^{(2)}]_{i,j}$.*

Next, we characterize the parameters of a restricted SEM, and then formalize the observations stated in Remark 1 in the next proposition.

Lemma 1 ([9]) *Corresponding to a subset $S \subseteq [p]$, denote the removed set of nodes by $U \triangleq [p] \setminus S$ and define $U_j \triangleq U \cap \mathsf{an}(j)$, for $j \in S$. We have*

$$\sigma_{S,j}^2 = \sigma_j^4 \left(\sigma_j^2 - B_{U_j,j}^\top [\Theta_{\mathsf{an}(j)}]_{U_j,U_j}^{-1} B_{U_j,j}\right)^{-1}, \tag{10}$$

$$[B_S]_{k,j} = \frac{\sigma_{S,j}^2}{\sigma_j^2}\left(B_{k,j} - B_{U_j,j}^\top [\Theta_{\mathsf{an}(j)}]_{U_j,U_j}^{-1} [\Theta_{\mathsf{an}(j)}]_{U_j,k}\right). \tag{11}$$

Proposition 1 *Denote the ancestors of $j \notin \mathcal{I}$ in $\mathcal{I}$ by $\mathsf{an}_\mathcal{I}(j)$. If a set S contains $\mathsf{an}_\mathcal{I}(j)$ and their parents $\mathsf{pa}(\mathsf{an}_\mathcal{I}(j))$, then $\sigma_{S,j}^{(1)} = \sigma_{S,j}^{(2)}$. If set S does not contain any descendant of j we have $[\Delta_{\Theta_S}]_{j,j} = 0$. Furthermore, for $i \in \mathcal{I}$ and any set S we have $[\Delta_{\Theta_S}]_{i,i} \neq 0$. Furthermore, if $[B_S]_{j,i} \neq 0$ in either model, then we have $[\Delta_{\Theta_S}]_{j,i} \neq 0$.*

Remark 2 *We repeatedly use the restricted SEM characterization in Lemma 1 with various strategic choices of subsets S in Algorithm 1 to eliminate the non-intervened nodes from S_Δ using the criterion of Proposition 1. In Step 1, to identify J_0 in (7), we set $S = \{j\}$ for each $j \in S_\Delta$. In Step 2, to*

identify J_0^k in (8) for each $k \in S_\Delta \setminus J_0$, we set $S = \{j, k\}$ for each $j \in J_0$. In Step 3, to process nodes in $\mathcal{A}_\ell$, we use subsets of the form $\mathcal{M}_\ell \cup A$ for subsets A in $\mathcal{A}_\ell$.

Theorem 1 (Consistency) *Given Assumption 1 and the population covariance matrices, Algorithm 1 is consistent in estimating intervention target set $\mathcal{I}$ under soft interventions with $\mathsf{P} = 1$.*

4.2 $\mathcal{I}$-Markov equivalence

Intervening on a DAG changes the conditional distributions of the intervened variables, and hence, reveals orientations of some edges that were previously undirected in observational CPDAG, resulting in the interventional CPDAG ($\mathcal{I}$-CPDAG). The DAGs that have the same $\mathcal{I}$-CPDAG under soft intervention $\mathcal{I}$ form the $\mathcal{I}$-Markov equivalence class ($\mathcal{I}$-MEC). This is shown and discussed next.

For a DAG $\mathcal{G}$ and the intervention set $\mathcal{I}$, additional $\mathcal{I}$-vertex ζ and $\mathcal{I}$-edges $\{\zeta \to i\}_{i \in \mathcal{I}}$ are added to form the interventional DAG ($\mathcal{I}$-DAG). Note that vertex ζ creates a new v-structure $\zeta - i - j$ for any non-intervened $j \in \mathsf{pa}(i)$. However, if j is also in $\mathcal{I}$, then $\mathcal{I}$-DAG also contains the $\zeta \to j$ edge, and there is no new v-structure that can orient the edge $i - j$.

We call the edges in $\mathcal{G}$ that are not directed in the original CPDAG but are directed in $\mathcal{I}$-CPDAG as $\mathcal{I}$-directed edges. In the *parent finder* step of Algorithm 1, we find the edge set $\{j \to i\}_{j \notin \mathcal{I}, i \in \mathcal{I}}$, and subsequently, obtain the $\mathcal{I}$-MEC. Therefore, we can use Algorithm 1 in conjunction with an observational algorithm to perform causal structure learning, and establish the following theorem.

Theorem 2 ($\mathcal{I}$-MEC) *Algorithm 1 consistently recovers non-intervened parents of an $i \in \mathcal{I}$ with $\mathsf{Q} = 1$ in population setting. This results modifies the original MEC, which can be obtained via any observational structure learning algorithm, into the $\mathcal{I}$-MEC.*

4.3 Sample complexity

In this subsection, we provide the finite-sample versions of Theorems 1 and 2. Our choice of PDE function, the ADMM-based method of [7], comes with finite-sample results when noise ϵ has a Gaussian distribution. The following theorem establishes the sample complexity of Algorithm 1 for estimating $\mathcal{I}$ and their non-intervened parents.

Theorem 3 (Sample complexity) *Let d denote the maximum degree of an intervened node and set $\Gamma \triangleq \Sigma_2 \otimes \Sigma_1$ and $\alpha \triangleq 1 - \max_{e \notin \text{supp}} |\Gamma_{e,\text{supp}} \Gamma^{-1}_{\text{supp},\text{supp}}|_1$. Accordingly, define*

$$M \triangleq \max\{\|\Sigma^{(1)}\|_\infty, \|\Sigma^{(2)}\|_\infty\}, \quad M_\Sigma \triangleq \max\{\|\Sigma^{(1)}\|_{1,\infty}, \|\Sigma^{(2)}\|_{1,\infty}\}, \quad M_\Gamma \triangleq \|\Gamma^{-1}_{\text{supp},\text{supp}}\|_{1,\infty}.$$

When $M_\Sigma, M_\Gamma < +\infty$ and $\alpha > 0$, with $n = O\left(\frac{d^4}{\varepsilon^2} \frac{\log p}{\delta}\right)$ samples Algorithm 1:

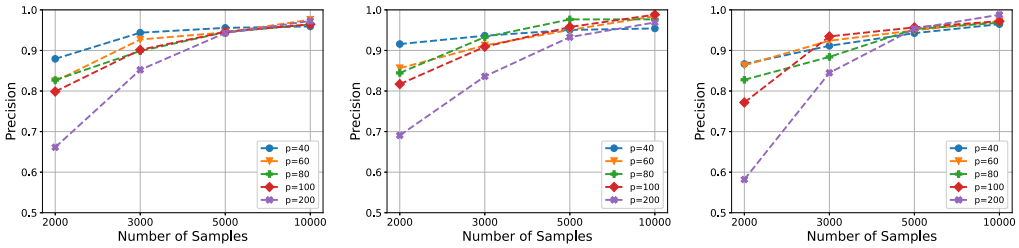
1. *identifies $\mathcal{I}$ with a probability at least $\geq 1 - \delta$;*
2. *identifies the non-intervened parents $\{j \to i\}_{j \notin \mathcal{I}, i \in \mathcal{I}}$ with a probability at least $\geq 1 - \delta$.*

Note that we have assumed M_Σ and M_Γ are bounded. This is necessary to avoid a linear scaling of the sample complexity in p.

5 Empirical results

5.1 Synthetic data - intervention recovery

We start by testing our algorithm for estimating intervention targets, i.e., the set $\mathcal{I}$. We generate 100 realizations of Erdős-Rényi [17] DAGs with expected neighborhood size $c = 1.5$, and $|\mathcal{I}| = 5$. We sample the entries of B, i.e., the edge weights, independently at random according to the uniform distribution on $[-1, -0.25] \cup [0.25, 1]$. The additive Gaussian noise terms have distribution $\mathcal{N}(0, I_p)$. We select the intervention set $\mathcal{I}$ by randomly selecting 5 nodes from $[p]$. We consider three different models to intervene on the nodes in $\mathcal{I}$: (i) *shift intervention* model in which mean of the noise ϵ_i is shifted from 0 to 1, (ii) *variance increase* model in which the variance of the noise ϵ_i is increased from 1 to 2, and (iii) *randomized intervention* model, in which $(B^{(2)})_{\mathsf{pa}(i),i} = 0$ and noise variance is increased from 1 to 1.5. All the simulations are run on a MacBook Pro with 2.7 GHz Dual-Core i5 core and 8 GB RAM.



(a) Shift intervention (b) Increased variance (c) Randomized intervention

Figure 1: Average precision of estimating intervention targets. Algorithm 1 reaches high precision with increasing number of samples even in large models for all settings.

We first run our algorithm by varying the graph size p and the number of samples. Figure 1 illustrates that our algorithm is able to recover intervention targets with high precision under all three intervention models. Having high precision is especially important in high dimensions, since a large false positive rate severely affects any downstream task such as structure learning. Recall rates are close to 1 and are omitted from the graph.

Next, we compare our results with that of UT-IGSP algorithm [15] for the shift intervention model. We note that UT-IGSP performs a greedy search to identify the sparsest permutation through CI tests, and it returns intervention targets as a by-product along with the learned causal structure. The computation time of UT-IGSP, hence, grows quickly with the size of the graph, reaching an average of 61.2 seconds for $p = 100$. Therefore, the complexity of structure learning and intervention target discovery in the high-dimensional regime is prohibitive. In contrast, our algorithm has comparable performance to UT-IGSP when $p = 100$ while requiring less than a second of runtime. Our algorithm's runtime scales gracefully when dimension is in the hundreds.

Table 1: $\mathcal{I}$ estimation for shift intervention model - 50 repetitions with 5000 samples - density 1.5

	UT-IGSP ([15])			Ours		
p	Precision	Recall	Time(s)	Precision	Recall	Time(s)
40	0.99 (0.04)	0.99 (0.04)	0.8	0.96 (0.09)	0.94 (0.09)	0.1
60	0.95 (0.07)	0.99 (0.05)	5.2	0.97 (0.07)	0.95 (0.10)	0.2
80	0.96 (0.08)	0.99 (0.04)	17.8	0.96 (0.08)	0.96 (0.10)	0.3
100	0.93 (0.11)	1 (0)	61.2	0.94 (0.09)	0.98 (0.07)	0.3

5.2 Synthetic data - causal structure learning

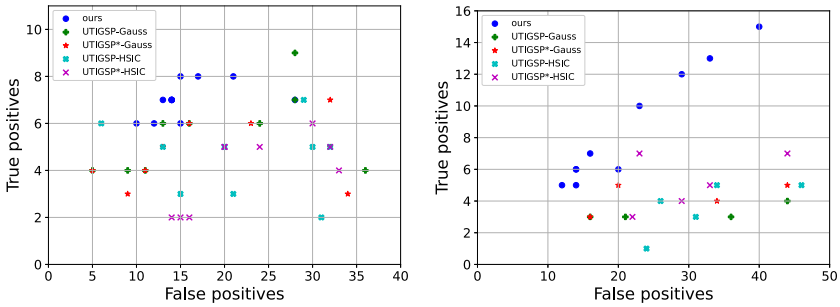
In Section 4.2 we have shown that Algorithm 1 recovers the intervention targets. It can be further used to refine observational MEC to an $\mathcal{I}$-MEC. Accordingly, we take the correct CPDAG of $\mathcal{G}^{(1)}$, then apply our algorithm's findings to obtain $\mathcal{I}$-CPDAG. We report the accuracy of additional edge orientations and in particular recovery of parents (if possible) of intervention targets in Appendix B.2.

5.3 Application to real data

We apply our algorithm on two real datasets with observational and interventional data to learn causal structures. When there exist multiple interventional environments, we apply our algorithm on them individually to estimate the intervened nodes and their parents. Subsequently, we combine the results from all environments in order to form the final estimated structure. There is a large number of interventional settings in which finding targets and their non-intervened parents by our algorithm yields a good estimate of the associated DAG, which we use for our evaluation.

We compare our results with those of algorithms in [15][1], UT-IGSP and UT-IGSP*, where the former works with partially known intervention targets and the latter does not require any target input. We use both parametric (Gaussian) and non-parametric (Hilbert-Schmidt independence criterion) CI

[1]The code and preprocessed real data are taken from https://github.com/csquires/utigsp for fair comparison, and CausalDAG package which is distributed under 3-Clause BSD licence is used.



(a) Sachs [18] dataset

(b) Dixit [20] dataset

Figure 2: ROC curves for directed edge recovery. Algorithm 1: (a) is more robust in smaller graphs; (b) handles larger graphs and multiple interventional settings successfully, and conforms to real datasets better than CI testers.

tests for UT-IGSP methods. We note that the non-parametric tests have significant runtimes. We note that our algorithm uses PDE at several stages, which calls for different λ regularization parameters. Namely, let us denote the parameters used for Step 1 and Step 3 by λ_1, Step 2 by λ_2 and *parent finder* of Step 3 by λ_3. Similarly, UT-IGSP needs a cut-off value α for CI tests. We run the algorithms with different values of these parameters to obtain the ROC curves.

Protein signaling data. We first consider the dataset of Sachs et al. [18] for the discovery of protein signaling network of 11 nodes. It consists of measurements of proteins and phospolipids under different interventional environments. In each environment, signaling nodes are inhibited or activated. Hence, these sites form intervention targets. The conventionally accepted ground truth has been updated over the years, and we compare with the recent version in [19], which consists of 16 edges. We follow the process of [15] and work with 1755 observational and 4091 interventional samples aggregated from 5 different interventional environments. In Fig. 2a, we report the results of running Algorithm 1 and UT-IGSP with various parameters.

Perturb-seq gene expression data. We analyze the performance of our algorithm on perturb-seq dataset by Dixit et al. [20]. The dataset consists of observational data and the interventional data from bone marrow-derived dendritic cells (BMDCs). A single gene has been targeted for deletion in each interventional environment. Similar to [20] we focused on 24 genes that are known to have regulatory effects, and we followed [11] to select interventional data from 8 gene deletions along with observational samples. We take the Fig. 4D of Dixit et al. [20] as the ground truth that has 34 edges among 24 nodes. We use 23 interventional settings for the targeted 8 gene deletions. In Fig. 2b, we plot the results of running Algorithm 1 and UT-IGSP with various parameters. In both real datasets, our algorithm achieves higher accuracy in recovering directed edges. The comparison with UT-IGSP is more striking in Fig. 2b, and shows that our ability to work with many interventional environments on a relatively larger graph. Furthermore, Fig. 2a shows that Algorithm 1 handles smaller graphs more robustly. These results illustrate that even though our algorithm is designed for linear models, on real datasets, it performs better than the current state-of-the-art methods that rely on CI tests.

6 Conclusion

In this paper, we have considered the problem of estimating intervention targets in linear structural equation models (SEMs) under soft interventions. We have proposed an algorithm that consistently identifies intervened nodes that can scale to larger graphs and have sample complexity guarantees in Gaussian linear SEMs. The algorithm can be used also to infer interventional Markov equivalence class (MEC) from the observational MEC. We have demonstrated comparable or better performance compared to the existing methods in a number of settings.

The limitation of our method is that it only applies to linear SEMs. The dataset in an application should be evaluated carefully to confirm whether the assumptions are satisfied. This avoids any adverse effects arising from wrong interpretations of cause-effect relationships. Extending the similar ideas for scalable and efficient intervention target estimation to the non-linear DAGs is an open question that we aim to address in future work.

References

[1] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, and J. D. Sander, "High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells," *Nature biotechnology*, vol. 31, no. 9, pp. 822–826, 2013.

[2] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y.-E. Chiew, I. Haviv, A. O. C. S. Group, D. Gertig, A. deFazio, and D. D. Bowtell, "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5198–5208, 2008.

[3] F. Eberhardt and R. Scheines, "Interventions and causal inference," *Philosophy of Science*, vol. 74, no. 5, pp. 981–995, 2007.

[4] Y. Wang, C. Squires, A. Belyaeva, and C. Uhler, "Direct estimation of differences in causal graphs," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2018, pp. 3770–3781.

[5] S. D. Zhao, T. T. Cai, and H. Li, "Direct estimation of differential networks," *Biometrika*, vol. 101, no. 2, pp. 253–268, 2014.

[6] H. Yuan, R. Xi, C. Chen, and M. Deng, "Differential network analysis via lasso penalized d-trace loss," *Biometrika*, vol. 104, no. 4, pp. 755–770, 2017.

[7] B. Jiang, X. Wang, and C. Leng, "A direct approach for sparse quadratic discriminant analysis," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1098–1134, 2018.

[8] Z. Tang, Z. Yu, and C. Wang, "A fast iterative algorithm for high-dimensional differential network," *Computational Statistics*, vol. 35, no. 1, pp. 95–109, 2020.

[9] A. Ghoshal and J. Honorio, "Direct estimation of difference between structural equation models in high dimensions," *arXiv preprint arXiv:1906.12024*, 2019.

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

[11] Y. Wang, L. Solus, K. Yang, and C. Uhler, "Permutation-based causal inference algorithms with interventions," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, Dec. 2017, pp. 5822–5831.

[12] D. Eaton and K. Murphy, "Exact Bayesian structure learning from uncertain interventions," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, Mar. 2007, pp. 107–114.

[13] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, B. Schölkopf, M. C. Mozer, C. Pal, and Y. Bengio, "Learning neural causal models from unknown interventions," *arXiv preprint arXiv:1910.01075*, 2019.

[14] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," in *Proc. Advances in Neural Information Processing Systems*, virtual, Dec. 2020, pp. 21 865–21 877.

[15] C. Squires, Y. Wang, and C. Uhler, "Permutation-based causal structure learning with unknown intervention targets," in *Proc. Conference on Uncertainty in Artificial Intelligence*, virtual, Aug. 2020, pp. 1039–1048.

[16] T. Verma and J. Pearl, "An algorithm for deciding if a set of observed independencies has a causal explanation," in *Proc. Conference on Uncertainty in Artificial Intelligence*, Stanford, CA, Jul. 1992, pp. 323–330.

[17] P. Erdös and A. Rényi, "On Random Graphs I," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

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

[19] R. O. Ness, K. Sachs, P. Mallick, and O. Vitek, "A Bayesian active learning experimental design for inferring signaling networks," in *Proc. Research in Computational Molecular Biology*, Hong Kong, China, May 2017, pp. 134–156.

[20] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev, "Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens," *Cell*, vol. 167, no. 7, pp. 1853–1866.e17, 2016.

[21] A. Ghoshal and J. Honorio, "Learning linear structural equation models in polynomial time and sample complexity," in *Proc. International Conference on Artificial Intelligence and Statistics*, Playa Blanca, Lanzarote, Canary Islands, Apr. 2018, pp. 1466–1475.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the last paragraph in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the last paragraph in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 for exact problem definition and Assumption 1 in Section 4 for assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the code in supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and the code in supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Standard deviation of the precision and recall rates of intervention recovery are given in Table 1 in Section 5.1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See first paragraph in Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.
   (b) Did you mention the license of the assets? [Yes] See the footnote in Section 5.3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code for our algorithm and simulations are provided in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]