Linear Causal Disentanglement via Interventions

Chandler Squires [* 1 2] **Anna Seigal** [* 1 3] **Salil Bhate** [1] **Caroline Uhler** [1 2]

Abstract

Causal disentanglement seeks a representation of data involving latent variables that are related via a causal model. A representation is identifiable if both the latent model and the transformation from latent to observed variables are unique. In this paper, we study observed variables that are a linear transformation of a linear latent causal model. Data from interventions are necessary for identifiability: if one latent variable is missing an intervention, we show that there exist distinct models that cannot be distinguished. Conversely, we show that a single intervention on each latent variable is sufficient for identifiability. Our proof uses a generalization of the RQ decomposition of a matrix that replaces the usual orthogonal and upper triangular conditions with analogues depending on a partial order on the rows of the matrix, with partial order determined by a latent causal model. We corroborate our theoretical results with a method for causal disentanglement. We show that the method accurately recovers a latent causal model on synthetic and semi-synthetic data and we illustrate a use case on a dataset of single-cell RNA sequencing measurements.

1. Introduction

The goal of representation learning is to find a description of data that is interpretable, useful for reasoning, and generalizable. Such a representation *disentangles* the data into conceptually distinct variables. Traditionally, conceptual distinctness of variables has meant statistical independence. This is the setting of independent component analysis (Comon, 1994). However, human reasoning often involves variables that are not statistically independent. For example, the presence of a sink and the presence of a mirror in an image. It is therefore natural to generalize conceptual distinctness to variables that are *causally autonomous*; i.e., interventions can be performed on each variable separately. This motivates *causal disentanglement* (Yang et al., 2021), the recovery of causally autonomous variables from data.

In this paper, we study the identifiability of causal disentanglement; i.e., its uniqueness. We adopt a generative perspective, as in (Bengio et al., 2013; Moran et al., 2021). We assume that the observed variables are generated in two steps. First, latent variables Z are sampled from a distribution $\mathbb{P}(Z)$. Then, the observed variables X are the image of the latent variables under a deterministic mixing function. We assume that the latent variables are generated according to a linear structural equation model and that the mixing function is an injective linear map. Recent work has studied identifiability of various settings in representation learning (Khemakhem et al., 2020; Ahuja et al., 2021). A common assumption for identifiability is that variables are observed across multiple contexts, each affecting the latent distribution $\mathbb{P}(Z)$ but not the mixing function. In our setup, each context is either an *intervention* on a latent variable, or is *observational*, i.e., has no interventions. We use the same terminology for interventions as Squires & Uhler (2022). From most to least general, a *soft* intervention on Z_i changes the dependency of Z_i on its direct causes, a *perfect* intervention removes this dependency but allows for stochasticity of Z_i, and a *do*-intervention sets Z_i to a deterministic value.

Our main result is that our linear causal disentanglement setup is identifiable if, in addition to an observational context, for each latent variable Z_i, there is a context where Z_i is the intervened variable under a perfect intervention; see Section 3.2. This is a *sufficient* condition for identifiability. Furthermore, we show that the condition of at least one intervention per latent node is *necessary* in the worst case: if some latent node is not intervened in any context, then there exist latent causal representations that are not identifiable; see Section 3.3. Our focus in this paper is on identifiability guarantees. Nonetheless, we convert our proofs into a method for causal disentanglement in the finite-sample setting. In Section 4, we apply the method to synthetic and semi-synthetic data and show that it recovers the generative model, and we compute a linear causal disentanglement on a single-cell RNA sequencing dataset.

---

[*]Equal contribution [1]Broad Institute of MIT and Harvard [2]Laboratory for Information and Decision Systems, MIT [3]School of Engineering and Applied Sciences, Harvard University. Correspondence to: Chandler Squires <csquires@mit.edu>.

1.1. Motivating Example

Consider two latent variables $Z = (Z_1, Z_2)$. Assume that $X = (X_1, X_2)$ is observed in two contexts $k \in \{0, 1\}$, that $X = GZ$ in both contexts, and that in context k,

$$Z = A_k Z + \Omega_k^{1/2} \varepsilon \quad \text{for} \quad \varepsilon \sim \mathcal{N}(0, I).$$

Let

$$A_0 = A_1 = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}, \quad \Omega_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\Omega_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}, \quad G = \begin{bmatrix} -2 & 2 \\ 2 & -1 \end{bmatrix}.$$

Context $k = 1$ is an intervention on Z_2 that changes its variance. The covariance of X in contexts $k = 0$ and $k = 1$ are, respectively,

$$\Sigma_0 = \begin{bmatrix} 20 & -16 \\ -16 & 13 \end{bmatrix} \quad \text{and} \quad \Sigma_1 = \begin{bmatrix} 8 & -7 \\ -7 & 25/4 \end{bmatrix},$$

since $\Sigma_k = G(I - A_k)^{-1} \Omega_k (I - A_k)^{-\top} G^\top$. However, the following parameters give the same covariance matrices:

$$\widehat{A}_0 = \widehat{A}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \widehat{\Omega}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\widehat{\Omega}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}, \quad \widehat{G} = \begin{bmatrix} -2 & 4 \\ 2 & -3 \end{bmatrix}.$$

The second set of parameters imply independence of Z_1 and Z_2, since $(\widehat{A}_0)_{1,2} = 0$, whereas the original parameters imply non-independence since $(A_0)_{1,2} \neq 0$. This non-identifiability holds for generic $A_0, A_1, \Omega_0, \Omega_1$, and G, where A_1 comes from an intervention on Z_2, see Section 3.3. This non-identifiability extends to any number of latent variables d: we show that, in the worst case, non-identifiability holds when fewer than $d + 1$ contexts are observed.

1.2. Related Work

The growing field of causal representation learning blends techniques from multiple lines of work. Chief among these are identifiable representation learning, causal structure learning, and latent DAG (directed acyclic graph) learning.

Identifiable representation learning. The identifiability of the linear independent component analysis (ICA) model was given in Comon (1994). Identifiability of a nonlinear ICA model is studied in (Hyvarinen et al., 2019; Khemakhem et al., 2020), in the presence of auxiliary variables. The ICA model imposes the stringent condition that the latent variables are independent (in the linear setting) or conditionally independent given the auxiliary variables (in the nonlinear setting). Recent works on identifiable representation learning (Ahuja et al., 2021; Zimmermann et al., 2021) introduce structure on the data generating process to circumvent the independence condition. However, they do not consider latent variables that are causally related.

Causal Structure Learning. Causal structure is identifiable up to an *equivalence class* that depends on the available interventional data (Verma & Pearl, 1990; Hauser & Bühlmann, 2012; Yang et al., 2018; Squires et al., 2020; Jaber et al., 2020). See Squires & Uhler (2022) for a recent review. A key line of work (Eberhardt et al., 2005; Hyttinen et al., 2013) characterizes the interventions necessary and sufficient to ensure that the causal structure is fully identifiable; i.e., that the equivalence class is of size one. In particular, Eberhardt et al. (2005) showed that $d - 1$ interventions are in the worst case necessary to fully identify a causal DAG model on d nodes. The current paper extends this line of work to DAG models over *latent* variables.

Learning latent DAG models. The task of learning a DAG over latent variables dates back at least to Silva et al. (2006). They introduced the notion of a *pure child*: an observed variable X_i with only one latent parent, such X_i is also called an *anchor* (Halpern et al., 2015; Saeed et al., 2020). The method of Silva et al. (2006) requires that all observed variables are pure children. Recent works relax this assumption by studying the linear non-Gaussian setting, where all latent and observed variables are linear functions of their parents plus independent non-Gaussian noise. For example, Cai et al. (2019) propose a method which learns a latent DAG under the assumption that each latent variable has at least two pure children. The pure child assumption can be extended to allow subsets of latent variables with the same observed children, as in Xie et al. (2020), which introduces the Generalized Independent Noise condition. This condition was used by Xie et al. (2022) to permit latent variables with no observed children; i.e., a hierarchical latent model.

Other works consider the discrete setting, requiring that the latent and observed variables are discrete (Halpern et al., 2015) or that the latent variables are discrete (Kivva et al., 2021). The paper Kivva et al. (2021) relaxes the pure child assumption, as follows. The children of node Z_i are the variables with a directed edge from Z_i. The *no twins* assumption says that the observed children of any two latent nodes are distinct sets. A similar assumption called *strong non-redundancy* appears in Adams et al. (2021), which considers models whose latent variables can be downstream of observed variables. See Appendix A. These works require sparsity in the map between latent and observed variables: they do not allow all observed variables to be children of all latent variables, which is the setting of the present paper.

A number of recent works discard the sparsity requirement. Ahuja et al. (2022a) and Brehmer et al. (2022) learn a latent DAG from paired counterfactual data. In contrast, we study *unpaired* data, which is more realistic in applications such as biology (Stark et al., 2020). To the best of our knowledge, only two works consider unpaired data without sparsity assumptions. Liu et al. (2022) study a linear Gaussian model

	Setting	Graphical Conditions	Identification Result
Silva et al. (2006)	**LNG**	All children pure	Identified up to Markov equivalence.
Halpern et al. (2015)	**Dis**	1 pure child per latent	Identified up to Markov equivalence.
Cai et al. (2019)	**LNG**	2 pure children per latent	Fully identified.
Xie et al. (2020)	**LNG**	2 pure children per latent	Fully identified.
Xie et al. (2022)	**LNG**	2 pure children per latent*	Fully identified.
Kivva et al. (2021)	**Dis**	No twins	Identified up to Markov equivalence.
Liu et al. (2022)	**LG**	None	Fully identified from $2|V|$ perfect interventions if $|E| \leq |V|$.
Ahuja et al. (2022b)	**Poly**	None	Fully identified from $|V|$ do-interventions.
This paper	**LNG** or **LG**	None	Fully identified from $|V|$ perfect interventions.

Table 1: **Settings from prior works on learning latent DAG models. LNG** is short for *linear non-Gaussian*, **LG** for *linear Gaussian*, **Dis** for **discrete**, and **Poly** for *polynomial mixing*. In *, pure children are allowed to be latent. The number of nodes and the number of edges in the latent graph are denoted $|V|$ and $|E|$, respectively.

over d latent variables, a nonlinear mixing function, and vector-valued contexts. Their identifiability result only applies to our setting if the latent graph has at most as many edges as nodes, see Appendix K. In that case, their result implies that $2d$ interventions suffice for identifiability. We strengthen this result, showing that (i) d interventions are sufficient and (ii) no restrictions on the latent graph are required. Moreover, we show that d interventions are, in the worst case, necessary. Such necessary conditions do not appear in prior work on identifying latent DAGs. Finally, contemporaneous work (Ahuja et al., 2022b) shows that a latent DAG is identifiable from the more restricted class of do-interventions, but allow non-linear relationships. See Table 1 for a summary of prior work.

2. Setup

We consider d latent variables $Z = (Z_1, \ldots, Z_d)$, generated according to a linear structural equation model. We index contexts by $k \in \{0\} \cup [K]$, where $[K] := \{1, \ldots, K\}$. The linear structural equation models in each context are related: context $k = 0$ is *observational* data, while contexts $k \in [K]$ are *interventional* data. We now state the assumptions for our model; see also Fig. 1.

Assumption 1.

(a) ***Linear latent model***: *Let $\mathcal{G}$ be a DAG with nodes ordered so that an edge $j \to i$ implies $j > i$. The latent variables Z follow a linear structural equation model: in context k, the latent variables Z satisfy*

$$Z = A_k Z + \Omega_k^{1/2}\varepsilon, \qquad Cov(\varepsilon) = I_d,$$

where $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix, $\Omega_k \in \mathbb{R}^{d \times d}$ is diagonal with positive entries, and $A_k \in \mathbb{R}^{d \times d}$ has $(A_k)_{ij} \neq 0$ if and only if there is an edge $j \to i$ in $\mathcal{G}$. That is, in context k,

$$Z = B_k^{-1}\varepsilon, \quad \text{where} \quad B_k = \Omega_k^{-1/2}(I_d - A_k). \quad (1)$$

(b) ***Generic single-node interventions:*** *For each $k \in [K]$, there exists $i_k \in \{1, \ldots, d\}$ such that*

$$B_k = B_0 + \mathbf{e}_{i_k}\mathbf{c}_k^\top,$$

further, $(B_k)^\top\mathbf{e}_{i_k}$ is not a multiple of $(B_0)^\top\mathbf{e}_{i_k}$ unless i_k has no parents in $\mathcal{G}$.

(c) ***Linear observations:*** *Fix $p \geq d$. There is a full rank matrix $G \in \mathbb{R}^{p \times d}$ such that $X = GZ$ in every context k. Let $H := G^\dagger$ denote its Moore-Penrose pseudoinverse. We set the entry of largest absolute value in each row of H to 1. If multiple entries in a row have same absolute value we set the leftmost entry to be positive.*

Our strongest results hold under one additional assumption.

Assumption 2. ***Perfect interventions:*** *For each $k \in [K]$, there exists $i_k \in \{1, \ldots, d\}$ such that*

$$B_k = B_0 + \mathbf{e}_{i_k}\mathbf{c}_k^\top,$$

where $\mathbf{c}_k = \lambda_k\mathbf{e}_{i_k} - B_0^\top\mathbf{e}_{i_k}$ for some $\lambda_k > 0$.

Remark 1 (The parts of Assumption 1 that hold without loss of generality). *Taking $Var(\varepsilon_i) = 1$ for all i holds without loss of generality, since scaling can be absorbed into the matrix Ω_k. A linear structural equation model is causally sufficient if $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$ for all $i \neq j$. Thus, for a causally sufficient linear structural equation model, we have $Cov(\varepsilon) = I_d$ in Assumption 1(a) without loss of generality. The ordering of nodes in Assumption 1(a) is also without loss of generality: a permutation of the latent nodes can be absorbed into G. Our ordering makes the matrices A_k upper triangular.*

The scaling of H in Assumption 1(c) is without loss of generality, as follows. If $\{B_k\}_{k=0}^K$ and H satisfy Assumption 1 then $X = (B_k H)^\dagger\varepsilon$. Consider the matrices $\{B_k\Lambda\}_{k=0}^K$ and $\Lambda^{-1}H$, for Λ diagonal with positive entries. Observe that $X' = (B_k\Lambda\Lambda^{-1}H)^\dagger\varepsilon$, has the same distribution as X in context k. The alternative matrices satisfy Assumption 1, except for the scaling condition on H. Assumption 1(c) therefore fixes the scaling indeterminacy of each node.
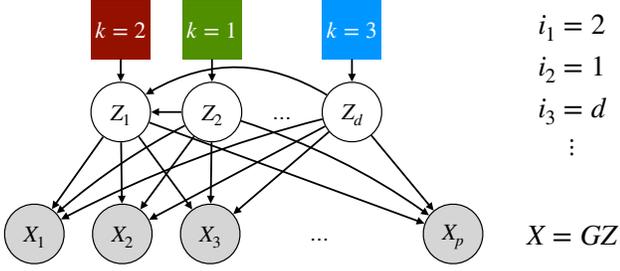
Figure 1: **The proposed setup.** The latent variables $Z = (Z_1, \ldots, Z_d)$ follows a linear DAG model, with contexts $k = 1, \ldots, K$ being single node interventions on targets $i_1, \ldots, i_K$. The observed variables $X = (X_1, \ldots, X_p)$ are an injective linear function of the latent variables $X = GZ$, where $G \in \mathbb{R}^{p \times d}$ does not vary across contexts.

The genericity condition in Assumption 1(b) automatically holds for perfect interventions. It fails to hold only for soft interventions that changes the variance but not the edge weights[1]. We show the importance of the genericity assumption for the identifiability of causal disentanglement in Appendix B.

We give an example of a setting in which Assumption 1 might apply. Suppose Z is the internal state of a cell (e.g., the concentrations of proteins, the locations of organelles, etc.) and that each context is an exposure to a different small molecule. Assumption 1(b) posits that each small molecule has a highly *selective* effect, modifying only one cellular mechanism. Assumption 2 posits that each small molecule completely disrupts the modified mechanism. While one does not expect all small molecules to be highly selective, one could filter based on selectivity.

In Appendix G, we describe a hypothesis test to test implications of Assumption 1(b), and show empirically that this test effectively determines model membership.

The covariance of X in context k is rank deficient when $d < p$, since $X = GZ$. We therefore define the precision matrix of X in context k to be the pseudoinverse of the covariance matrix, $\Theta_k := \mathbb{E}[XX^\top]^\dagger$. Then

$$\Theta_k = H^\top B_k^\top B_k H, \tag{2}$$

by Prop. 6 in Appendix C, since $\mathrm{Cov}_k(Z) = (B_k^\top B_k)^{-1}$.

We consider an unknown latent DAG. Each candidate DAG has unknown weights on its edges, unknown variances on its nodes, unknown new weights under each intervention, and an unknown mixing map to the observed variables. That is, our goal is to decompose the precision matrices $\{\Theta_k\}_{k=0}^K$ in Equation (2) to recover H and $\{B_k\}_{k=0}^K$.

We recall some graph theoretic notions. The parents of node i are $\mathrm{pa}_{\mathcal{G}}(i) = \{j \mid j \to i \text{ in } \mathcal{G}\}$, and we define $\overline{\mathrm{pa}}_{\mathcal{G}}(i) :=$

---

[1]i.e., $(\Omega_k)_{i_k, i_k} \neq (\Omega_0)_{i_k, i_k}$ and $(A_k)^\top \mathbf{e}_{i_k} = (A_0)^\top \mathbf{e}_{i_k}$

$\mathrm{pa}_{\mathcal{G}}(i) \cup \{i\}$. Similarly, $\mathrm{an}_{\mathcal{G}}(i)$ denotes the *ancestors* of i in $\mathcal{G}$, the vertices j with a directed path from j to i. We define $\overline{\mathrm{an}}_{\mathcal{G}}(i) := \mathrm{an}_{\mathcal{G}}(i) \cup \{i\}$ and $\overline{\mathrm{an}}_{\mathcal{G}}(\mathcal{I}) := \cup_{i \in \mathcal{I}} \overline{\mathrm{an}}_{\mathcal{G}}(i)$. The source nodes of $\mathcal{G}$ are the nodes i with $\mathrm{pa}_{\mathcal{G}}(i) = \varnothing$. We drop the subscript $\mathcal{G}$ when the graph is clear from context.

The *transitive closure* of $\mathcal{G}$, denoted $\overline{\mathcal{G}}$, is the DAG with $\mathrm{pa}_{\overline{\mathcal{G}}}(i) = \mathrm{an}_{\mathcal{G}}(i)$. Given a DAG $\mathcal{G}$, define the partial order $\prec_{\mathcal{G}}$ to be $i \prec_{\mathcal{G}} j$ if and only if $j \in \mathrm{an}_{\mathcal{G}}(i)$. Thus, the transitive closure is the graph with $j \to i$ whenever $i \prec_{\mathcal{G}} j$.

To decompose the precision matrices in Equation (2), we introduce a matrix decomposition defined on a partial order. Recall that the RQ decomposition writes $H \in \mathbb{R}^{d \times p}$ as $H = RQ$ for an upper triangular $R \in \mathbb{R}^{d \times d}$ and orthogonal $Q \in \mathbb{R}^{d \times p}$. We generalize the RQ decomposition[2].

Definition 1 (The partial order RQ decomposition)**.** *Given a partial order $\prec$, the partial order RQ decomposition writes $H \in \mathbb{R}^{d \times p}$ as $H = RQ$, where $R \in \mathbb{R}^{d \times d}$ satisfies $R_{ii} \geq 0$ and $R_{ij} = 0$ unless $i \preceq j$, and where $\mathbf{q}_i$, the i-th row of $Q \in \mathbb{R}^{d \times p}$, is norm one and orthogonal to $\langle \mathbf{q}_j : i \prec j \rangle$.*

We recover the usual reduced RQ decomposition (Trefethen & Bau III, 1997) when $\prec$ is the total order $1 < \cdots < d$. We construct the partial order RQ decomposition in Appendix D. Finally, given a positive definite matrix $M \in \mathbb{R}^{d \times d}$, the *Cholesky factor* $U \in \mathbb{R}^{d \times d}$, denoted CHOLESKY($M$), is the unique upper triangular matrix with positive diagonal such that $M = U^\top U$.

3. Identifiability of Causal Disentanglement

We establish the sufficiency and worst case necessity of one intervention per latent node for identifiability of our causal disentanglement problem. The following result describes recovery of $\mathcal{G}$. Later, we discuss identifiability of the parameters in our setup. Since the labeling of latent nodes is unimportant, $\mathcal{G}$ is recovered if it is found up to relabeling.

Theorem 1. *Assume the setup in Assumption 1 with d latent variables. Then d interventions are sufficient and, in the worst case, necessary to recover $\overline{\mathcal{G}}$ from $\{\Theta_k\}_{k \in K}$. If Assumption 2 also holds, then d interventions are sufficient and, in the worst case, necessary to recover $\mathcal{G}$ from $\{\Theta_k\}_{k \in K}$.*

3.1. Preliminaries

We note the following basic fact, where $\mathbf{v}^{\otimes 2} := \mathbf{v}\mathbf{v}^\top$:

Fact 1. *Let $B \in \mathbb{R}^{d \times d}$. Then $B^\top B = \sum_{i=1}^d (B^\top \mathbf{e}_i)^{\otimes 2}$.*

We give a proof in Appendix F. This fact gives the key identity that drives our identifiability results.

---

[2]The RQ decomposition is used here (rather than the more familiar QR decomposition) because it gives an expression for the *rows* of H, which each correspond to one latent variable.

Proposition 1. *Consider the setup in Assumption 1. Then, for any $k \in [K]$,*

$$\Theta_k - \Theta_0 = (H^\top B_k^\top \mathbf{e}_{i_k})^{\otimes 2} - (H^\top B_0^\top \mathbf{e}_{i_k})^{\otimes 2}.$$

In particular the rank of the difference $\Theta_k - \Theta_0$ is 1 if i_k is a source node in $\mathcal{G}$, and 2 otherwise.

Proof. By Assumption 1, $B_k^\top \mathbf{e}_i = B_0^\top \mathbf{e}_i$ for all $i \neq i_k$. Using Fact 1, we have

$$B_k^\top B_k - B_0^\top B_0 = (B_k^\top \mathbf{e}_{i_k})^{\otimes 2} - (B_0^\top \mathbf{e}_{i_k})^{\otimes 2}.$$

Recall from Equation (2) that $\Theta_k = H^\top B_k^\top B_k H$. The result follows from left-multiplying both sides of the displayed equation by $H^\top$ and right-multiplying by H. This shows that $\mathrm{rank}(\Theta_k - \Theta_0) \leq 2$. For a source node, both vectors $B_k^\top \mathbf{e}_{i_k}$ and $B_0^\top \mathbf{e}_{i_k}$ have just one entry non-zero and $\mathrm{rank}(\Theta_k - \Theta_0) = 1$. Otherwise, the vectors have more than one entry non-zero and, by the genericity condition in Assumption 1(b), the difference $\Theta_k - \Theta_0$ has rank two. □

We can reduce a more general causal disentanglement problem to our setting, as we explain in Appendix F. First, we can count the latent dimension, since $\mathrm{rank}(\Theta_k) = d$ for any k. Second, we can identify which environments correspond to interventions on the same intervention target, see Prop. 9. Finally, we can identify which environment is observational using rank constraints, see Prop. 10. Thus, we assume without loss of generality that d is known, that the observational environment is known, and that each node is only intervened on in one context.

3.2. Sufficiency

We define $S(\mathcal{G})$ to be the set of permutations on d letters such that $\sigma(j) > \sigma(i)$ for all edges $j \to i$. For example, if $\mathcal{G}$ is a complete graph then $S(\mathcal{G})$ contains only the identity. If $\mathcal{G}$ has no edges then $S(\mathcal{G})$ is the group of permutations on d letters. The permutation matrix corresponding to permutation σ is $P_\sigma \in \mathbb{R}^{d \times d}$ with $(P_\sigma)_{ij} = \mathbb{1}_{\{i = \sigma(j)\}}$. Our main sufficiency result is the following.

Theorem 2. *Assume the set-up in Assumptions 1 and 2 with one intervention on each latent node. Then the graph $\mathcal{G}$, the intervention targets i_k, and the parameters are identifiable up to $S(\mathcal{G})$: given a solution $(B_0, \ldots, B_K, H)$, the set of solutions is $\{(P_\sigma B_0 P_\sigma^\top, \ldots, P_\sigma B_K P_\sigma^\top, P_\sigma H) : \sigma \in S(\mathcal{G})\}$.*

Theorem 2 says that solutions to the causal disentanglement problem are unique up to permutations of the latent nodes that preserve the property that $j \to i$ implies $j > i$. First, we verify that each permutation in $S(\mathcal{G})$ gives a solution.

Proposition 2. *Assume the set-up in Assumption 1. Given a solution $(B_0, \ldots, B_K, H)$ to Equation (2) for $k \in \{0\} \cup [K]$, the matrices $(P_\sigma B_0 P_\sigma^\top, \ldots, P_\sigma B_K P_\sigma^\top, P_\sigma H)$ are a valid solution whenever $\sigma \in S(\mathcal{G})$.*

Algorithm 1 ID-ANCESTORS

1: **Input:** $\Theta_k, \Theta_0, \{\widehat{\mathbf{q}}_i\}_{i \in \mathcal{I}}$
2: **Output:** Vector $\widehat{\mathbf{q}}_k$, ancestor set $\mathcal{A}$
3: Let $\mathcal{A} = \mathcal{I}$
4: **for** $i \in \mathcal{I}$ **do**
5:      Let $W_{\neg i} = \langle \widehat{\mathbf{q}}_i : j \in \mathcal{I} \setminus \{i\} \rangle$
6:      Let $V_{\neg i} = \mathrm{proj}_{W_i^\perp} \mathrm{rowspan}(\Theta_k - \Theta_0)$
7:      If $\dim(V_{\neg i}) = 1$, let $\mathcal{A} = \mathcal{A} \setminus \{i\}$
8: **end for**
9: Let $W = \langle \widehat{\mathbf{q}}_a : a \in \mathcal{A} \rangle$
10: Let $V = \mathrm{proj}_{W^\perp} \mathrm{rowspan}(\Theta_k - \Theta_0)$
11: Take $\widehat{\mathbf{q}}_k$ with first nonzero entry positive and $\|\widehat{\mathbf{q}}_k\|_2 = 1$, such that $\langle \widehat{\mathbf{q}}_k \rangle = V$
12: **return** $\widehat{\mathbf{q}}_k, \mathcal{A}$

Algorithm 2 ID-PARTIALORDER

1: **Input:** Precision matrices $(\Theta_0, \Theta_1, \ldots, \Theta_K)$, rank d
2: **Output:** Factor $\widehat{Q}$, partial order $\prec$
3: Let $\mathcal{I}_0 = \{\}, \widehat{Q} = \mathbf{0}_{d \times d}$
4: **for** $t = 1, \ldots K$ **do**
5:      Let $W_t = \langle \widehat{\mathbf{q}}_i : i \in \mathcal{I}_{t-1} \rangle$
6:      Let $V_k = \mathrm{proj}_{W_t^\perp} \mathrm{rowspan}(\Theta_k - \Theta_0)$ for $k \notin \mathcal{I}_{t-1}$
7:      Pick k such that $\dim(V_k) = 1$
8:      Let $\widehat{\mathbf{q}}_k, \mathcal{A} = $ ID-ANCESTORS$(\Theta_k, \Theta_0, \{\widehat{\mathbf{q}}_i\}_{i \in \mathcal{I}_{t-1}})$
9:      Add $a' \succ k$ for any $a' \succeq a, a \in \mathcal{A}$
10:      Let $\mathcal{I}_t = \mathcal{I}_{t-1} \cup \{k\}, \widehat{Q}_t = [\widehat{\mathbf{q}}_k; \widehat{Q}_{t-1}]$
11: **end for**
12: **return** $\widehat{Q}, \prec$

Proof. Let $\{B_k\}_{k=0}^K$ and H satisfy Assumption 1 and Equation (2). Define $B_k^{(\sigma)} = P_\sigma B_k P_\sigma^\top$ and $H^{(\sigma)} = P_\sigma H$ for $\sigma \in S(\mathcal{G})$. Then $\Theta_k = H^{(\sigma)\top} B_k^{(\sigma)\top} B_k^{(\sigma)} H^{(\sigma)}$. The matrices $B_k^{(\sigma)}$ are upper triangular, as follows. For all $i, j \in [p]$, we have $(B_k^{(\sigma)})_{\sigma(i),\sigma(j)} = (B_k)_{ij}$. Hence $B_k^{(\sigma)}$ is upper triangular when $(B_k)_{ij} = 0$ for all $i, j \in [p]$ with $\sigma(i) > \sigma(j)$. This holds since $\sigma \in S(\mathcal{G})$. Moreover, these matrices also satisfy Assumption 1(b) with the intervention target $\sigma(i_k)$ in context k. Finally, $H^{(\sigma)}$ satisfies Assumption 1(c), since we just permute the rows of H. □

We give a constructive proof of Theorem 2 via an algorithm to recover H and $\{B_k\}_{k=0}^K$ from $\{\Theta_k\}_{k=0}^K$ [3]. The computational complexity of the algorithm is given in Appendix H. The bulk of the algorithm is devoted to recovering H. First, we recover the partial order $\prec_\mathcal{G}$ (i.e., the DAG $\overline{\mathcal{G}}$), together with the matrix Q from a partial order RQ decomposition

---

[3] We only use the second moment of X. We do not use the first moment since we assume $\mathbb{E}[\varepsilon] = 0$, and we do not use higher moments since in the worst case (Gaussian noise), they contain no additional information.

of H, up to signs and permutations of rows, in Algorithm 2. The subroutine Algorithm 1 recovers the ancestors of a node and its corresponding row of Q. Then we recover R in Algorithm 3. Having recovered H, the matrices $\{B_k\}_{k=0}^{K}$ are found via the Cholesky decomposition.

We show that Algorithm 2 returns Q from the partial order RQ decomposition of H, up to a permutation $\sigma \in S(\mathcal{G})$ and a matrix in Sig_d, the $d \times d$ diagonal matrices with diagonal entries ± 1.

Proposition 3. *Assume the setup in Assumption 1, and that every latent node is intervened on; i.e, $\{i_k\}_{k=1}^{K} = [d]$. Let $(\widehat{Q}, \prec)$ be the output of Algorithm 2. Then $\prec$ is the partial order $\prec_{\mathcal{G}}$. Moreover, let $H = RQ$ be the partial order RQ decomposition of H for the partial order $\prec_{\mathcal{G}}$. Then $\widehat{Q} = SP_\sigma Q$ for some $\sigma \in S(\mathcal{G})$ and $S \in \text{Sig}_d$.*

The following lemma relates the partial order of $\mathcal{G}$ to the linear spaces $\text{rowspan}(\Theta_k - \Theta_0)$.

Lemma 1. *Assume the setup in Assumption 1. Let $H = RQ$ be a partial order RQ decomposition of H. Let i_k be the intervention target of context k and let $\mathcal{I} \subseteq [d]$. Then*

(a) $\text{rowspan}(\Theta_k - \Theta_0) \subseteq \langle \mathbf{h}_i : i \in \mathcal{I} \rangle$ if and only if $\overline{pa}(i_k) \subseteq \mathcal{I}$,

(b) $\text{rowspan}(\Theta_k - \Theta_0) \subseteq \langle \mathbf{q}_i : i \in \overline{an}(i_k) \rangle$, and

(c) $\text{rowspan}(\Theta_k - \Theta_0) \not\subseteq \langle \mathbf{q}_i : i \in \mathcal{I} \rangle$ if $\overline{pa}(i_k) \not\subseteq \mathcal{I}$.

Proof. Let $\mathbf{h}_i$ denote the ith row of H, and let $\boldsymbol{\alpha}_{j,i_k} := \sum_{i \in \overline{pa}(i_k)} (B_j)_{i_k,i} \mathbf{h}_i$. By Prop. 1, we have

$$\text{rowspan}(\Theta_k - \Theta_0) = \langle \boldsymbol{\alpha}_{k,i_k}, \boldsymbol{\alpha}_{0,i_k} \rangle. \tag{3}$$

Equation (3) shows that $\text{rowspan}(\Theta_k - \Theta_0) \subset \langle \mathbf{h}_i : i \in \overline{pa}(i_k) \rangle$. This linear space is contained in $\langle \mathbf{h}_i : i \in \mathcal{I} \rangle$ whenever $\overline{pa}(i_k) \subseteq \mathcal{I}$. Conversely, assume there exists $j \in \overline{pa}(i_k)$ with $j \notin \mathcal{I}$. Then, containment of $\text{rowspan}(\Theta_k - \Theta_0)$ in $\langle \mathbf{h}_i : i \in \mathcal{I} \rangle$ cannot hold: containment implies $(B_0)_{i_k,j} \mathbf{h}_j \in \langle \mathbf{h}_i : i \in [p] \setminus \{j\} \rangle$, a contradiction since H is full row rank and $(B_0)_{i_k,j} \neq 0$. Hence (a) holds.

By definition of the partial order RQ decomposition, we have $\mathbf{h}_i \in \langle \mathbf{q}_j \mid j \in \overline{an}(i) \rangle$. Thus, by transitivity of the ancestorship relation, (b) holds. Conversely, assume there exists $j \in \overline{pa}(i_k)$ with $j \notin \mathcal{I}$. Then $\text{rowspan}(\Theta_k - \Theta_0) \subseteq \langle \mathbf{q}_i : i \in \mathcal{I} \rangle$ implies that

$$\mathbf{h}_j \in \langle \mathbf{q}_i : i \in \mathcal{I} \rangle + \langle \mathbf{h}_i : i \in \overline{pa}(i_k) \setminus \{j\} \rangle,$$

since $(B_0)_{i_k,j} \neq 0$. We partition $\mathcal{I}$ into the descendants and non-descendants of j: let $\mathcal{I}_d := \{i \in \mathcal{I} : j \in \text{an}(i)\}$ and let $\mathcal{I}_{nd} := \{i \in \mathcal{I} : j \notin \text{an}(i)\}$. By definition of the partial order RQ decomposition, we have $\mathbf{q}_i \perp \mathbf{h}_j$ whenever $j \in \text{an}(i)$. Thus $\mathbf{h}_j \in \langle \mathbf{q}_i : i \in \mathcal{I}_{nd} \rangle + \langle \mathbf{h}_i : i \in$

Algorithm 3 ITERATIVEDIFFERENCEPROJECTION

1: **Input:** Precision matrices $(\Theta_0, \Theta_1, \ldots, \Theta_K)$
2: **Output:** $\widehat{H}, (\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_K)$
3: Let $d = \text{rank}(\Theta_0)$
4: Let $\widehat{Q}, \prec = \text{ID-PARTIALORDER}((\Theta_0, \Theta_1, \ldots, \Theta_K), d)$
5: Let $\widehat{C}_k = \text{CHOLESKY}((\widehat{Q}^\dagger)^\top \Theta_k \widehat{Q}^\dagger)$ for $k = 0, \ldots, K$
6: Let $\widehat{R} = I_d$
7: **for** $k = 1, \ldots, K$ **do**
8:     Let $\widehat{D}_k = \widehat{C}_k - \widehat{C}_0$
9:     Let $\widehat{i}_k$ be index of the only nonzero row of $\widehat{D}_k$
10:     Let $\widehat{R}_{\widehat{i}_k} = (\widehat{D}_k)_{\widehat{i}_k} + (\widehat{C}_0)_{\widehat{i}_k}$
11: **end for**
12: Let $\widehat{H}' = \widehat{R}\widehat{Q}$
13: Let $\widehat{H} = \widehat{\Lambda}\widehat{H}'$, for $\widehat{\Lambda}$ diagonal such that $\widehat{H}$ satisfies the conditions on H in Assumption 1(c)
14: Let $\widehat{B}_0 = \text{CHOLESKY}((\widehat{H}^\dagger)^\top \Theta_0 \widehat{H}^\dagger)$
15: Let $\widehat{B}_k = \widehat{B}_0 + \mathbf{e}_{\widehat{i}_k} \left(|\widehat{\Lambda}_{\widehat{i}_k, \widehat{i}_k}| \mathbf{e}_{\widehat{i}_k} - \widehat{B}_0^\top \mathbf{e}_{\widehat{i}_k} \right)^\top$ for $k = 1, \ldots, K$
16: **return** $\widehat{H}, (\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_K)$

$\overline{pa}(i_k) \setminus \{j\}\rangle$. Inverting the partial order RQ decomposition gives $\mathbf{q}_i \in \langle \mathbf{h}_{i'} : i' \in \overline{an}(i) \rangle$. Hence

$$\mathbf{h}_j \in \langle \mathbf{h}_i : i \in \overline{an}(\mathcal{I}_{nd}) \cup (\overline{pa}(i_k) \setminus \{j\}) \rangle,$$

a contradiction, since $j \notin \overline{an}(\mathcal{I}_{nd})$ and H is full rank. $\square$

Proof of Prop. 3. Assume that $\widehat{Q}_{t-1}$ is the last $t-1$ rows of $SP_\sigma Q$, for some $\sigma \in S(\mathcal{G})$ and $S \in \text{Sig}_d$. Then $W_{t-1} = \langle \widehat{\mathbf{q}}_i : i \in \mathcal{I}_{t-1} \rangle$. At step t, we pick k such that $V_k = \text{proj}_{W_t^\perp} \text{rowspan}(\Theta_k - \Theta_0)$ has dimension one. Lemma 1 implies that such k are those with $\text{pa}(i_k) \subseteq \mathcal{I}_{t-1}$. Algorithm 1 returns a set $\mathcal{A}$ with $\text{pa}(i_k) \subseteq \mathcal{A} \subseteq \text{ang}(i_k)$. Thus Line 2 adds the relation $k \prec a'$ if and only if $a' \in \text{an}_{\mathcal{G}}(i_k)$. Hence $\prec$ is the partial order $\prec_{\mathcal{G}}$. Line 2 picks $\widehat{\mathbf{q}}_k$ orthogonal to $\{\mathbf{q}_i : i \succ_{\mathcal{G}} i_k\}$, such that $\text{rowspan}(\widehat{Q}_t)$ contains $\text{rowspan}(\Theta_k - \Theta_0)$. By Lemma 1 and Definition 1, this is $\pm \mathbf{q}_{i_k}$. Thus, $\widehat{Q}_t$ is equal to the last t rows of $S'P_{\sigma'}Q$, for some $\sigma' \in S(\mathcal{G})$ and $S' \in \text{Sig}_d$. Repeating for $t = 1, \ldots, K$ gives the result. $\square$

We prove Theorem 2 by proving the following result, which is its constructive analogue.

Theorem 3. *Assume the setup in Assumptions 1 and 2, and that every latent node is intervened; i.e., $\{i_k\}_{k=1}^{K} = [d]$. Let $\widehat{H}$ and $\{\widehat{B}_k\}_{k=0}^{K}$ be the output of Algorithm 3. Then $\widehat{H} = P_\sigma H$ and $\widehat{B}_k = P_\sigma B_k P_\sigma^\top$ for all k, for some $\sigma \in S(\mathcal{G})$.*

Proof. Let $H = RQ$. Then $\Theta_k = Q^\top R^\top B_k^\top B_k RQ$, by Equation (2), and $\widehat{Q} = SP_\sigma Q$ for some $\sigma \in S(\mathcal{G})$ and

$S \in \mathrm{Sig}_d$, by Prop. 3. Hence $(\widehat{Q}^\dagger)^\top \Theta_k \widehat{Q}^\dagger$ equals

$$S \left(P_\sigma R^\top P_\sigma^\top \right) \left(P_\sigma B_k^\top P_\sigma^\top \right) \left(P_\sigma B_k P_\sigma^\top \right) \left(P_\sigma R P_\sigma^\top \right) S.$$

Let $C_k^{(\sigma)} = S(P_\sigma B_k P_\sigma^\top)(P_\sigma R P_\sigma^\top)S$. The matrix $C_k^{(\sigma)}$ is upper triangular, since it is a product of four upper triangular matrices, by the definition of $S(\mathcal{G})$ and of the partial order RQ decomposition, where we use that $i \prec_{\mathcal{G}} j$ implies $i < j$. Moreover, $C_k^{(\sigma)}$ has positive diagonal, since the matrices R and B_k have positive diagonal, by Definition 1 and Assumption 1(a) respectively. Hence $\widehat{C}_k := C_k^{(\sigma)}$ is the Cholesky factor of $(\widehat{Q}^\dagger)^\top \Theta_k \widehat{Q}^\dagger$. The differences $\widehat{D}_k := \widehat{C}_k - \widehat{C}_0$ equal $S P_\sigma (B_k - B_0) R P_\sigma^\top S = S P_\sigma \mathbf{e}_{i_k} \mathbf{c}_k^\top R P_\sigma^\top S$, by Assumption 1(b). The intervention target $\sigma(i_k)$ is the only nonzero row of $\widehat{D}_k$, i.e., $\widehat{i}_k = \sigma(i_k)$. Observe that $(\widehat{D}_k)_{\widehat{i}_k} + (\widehat{C}_0)_{\widehat{i}_k} = S_{\sigma(i_k),\sigma(i_k)} \lambda_k (R P_\sigma^\top)_{\sigma(i_k)}$. Thus, we have recovered $\widehat{R} = \widehat{\Lambda} P_\sigma R P_\sigma^\top$ for $\widehat{\Lambda}$ diagonal such that $\widehat{\Lambda}_{\widehat{i}_k,\widehat{i}_k} = \pm \lambda_k$. This gives $\widehat{H}' = \widehat{\Lambda} P_\sigma R P_\sigma^\top P_\sigma Q = \widehat{\Lambda} P_\sigma H$. The scaling in Line 3 recovers $\widehat{H} = P_\sigma H$ and $\widehat{\Lambda}$. We have $(\widehat{H}^\dagger)^\top \Theta_0 \widehat{H}^\dagger = P_\sigma B_0^\top P_\sigma^\top P_\sigma B_0 P_\sigma^\top$, where $P_\sigma B_0 P_\sigma^\top$ is upper triangular, and thus we recover $\widehat{B}_0 = P_\sigma B_0 P_\sigma^\top$ from the Cholesky decomposition. Finally, since $|\widehat{\Lambda}_{\widehat{i}_k,\widehat{i}_k}| = \lambda_k$, Line 3 gives us $\widehat{B}_k = P_\sigma B_k P_\sigma^T$. □

Theorem 3 requires Assumption 2, see Appendix J. In Appendix K, we compare our identifiability condition to that of Liu et al. (2022). We show that Liu et al. (2022) requires that the latent graph has fewer than d edges. In contrast, our condition imposes no constraints on the latent graph.

3.3. Worst-case Necessity

We show that one intervention per latent node is *necessary* for identifiability of our setup, in the worst case. It is clear that observational data does not suffice for identifiability: from just observational data, we always have a solution with *independent* latent variables, as follows. Let $\widehat{H} = \Lambda B_0 H$ and $\widehat{B}_0 = \Lambda^{-1}$, for Λ a diagonal matrix with positive entries such that Assumption 1(c) holds. Then $\widehat{B}_0 \widehat{H} = B_0 H$, i.e. $\widehat{B}_0, \widehat{H}$ solve the causal disentanglement problem. The new solution has independent latent nodes, since $\widehat{B}_0$ is diagonal.

The next result, which follows from prior work in causal structure learning, says that $d - 1$ interventions are required in the worst case, for a fully observed model. This is the special case of our setup where H is a permutation matrix.

Proposition 4. *Assume the setup in Assumptions 1 and 2, with H a permutation matrix. Let $K < d - 1$ with all intervention targets distinct. Then, in the worst case over intervention targets $\{i_k\}_{k=1}^K$, B_0 and H are not identifiable.*

Proof. In the linear Gaussian setting with unknown-target interventions, the structure of a DAG is only identifiable up to its *interventional Markov equivalence class* (MEC),

see e.g. Proposition 3.3(ii) of Castelletti & Peluso (2022). For single-node interventions, $d - 1$ interventions are in the worst case necessary to ensure that the interventional MEC has size one, by Theorem 3.7 of Eberhardt et al. (2005). □

We show that d interventions are necessary, in the worst case, when H is a general matrix. The proof is in Appendix I.

Proposition 5. *Assume the setup in Assumptions 1 and 2, with $K < d$. Then there exist H and $\{B_k\}_{k=0}^K$ and a set of K distinct intervention targets such that (i) H and B_k are not identifiable up to $S(\mathcal{G})$ and (ii) $\prec_{\mathcal{G}}$ is not identifiable.*

Example 1. *Proposition 5 generalizes our motivating example from Section 1. Fix $H \in \mathbb{R}^{2\times2}$ with entries H_{ij}, and fix upper triangular $B_0, B_1 \in \mathbb{R}^{2\times2}$ with entries $(B_0)_{ij}$ and $(B_1)_{ij}$, respectively. Assume $i_1 = 2$; i.e., $(B_0)_{11} = (B_1)_{11}$ and $(B_0)_{12} = (B_1)_{12}$. Let*

$$\widehat{B}_0 = \begin{bmatrix} 1 & 0 \\ 0 & (B_0)_{22} \end{bmatrix}, \quad \widehat{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & (B_1)_{22} \end{bmatrix},$$

$$\widehat{H} = \begin{bmatrix} (B_0)_{11} H_{11} + (B_0)_{12} H_{21} & (B_0)_{11} H_{12} + (B_0)_{12} H_{22} \\ H_{21} & H_{22} \end{bmatrix}.$$

Then for $k \in \{0, 1\}$, we have $\widehat{B}_k \widehat{H} = B_k H$, both equal to

$$\begin{bmatrix} (B_0)_{11} H_{11} + (B_0)_{12} H_{21} & (B_0)_{11} H_{12} + (B_0)_{12} H_{22} \\ (B_k)_{22} H_{21} & (B_k)_{22} H_{22} \end{bmatrix}$$

Remark 2. *Prop. 5 pertains to the worst case over intervention targets. It is natural to ask if there exists a better choice of K intervention targets, for $K < d$, such that H and $\{B_k\}_{k=0}^K$ are identifiable. For example, when $d = 2$, consider $\mathcal{G} = \{2 \to 1\}$, with an intervention on Z_1; i.e., $i_1 = 1$. Then $\mathrm{rowspan}(\Theta_1 - \Theta_0) \subseteq \langle \mathbf{h}_i : i \in \mathcal{I} \rangle$ if and only if $\mathcal{I} = \{1, 2\}$, by Lemma 1(a). Thus, $\Theta_1 - \Theta_0$ is rank 2, and we can detect that $i_1 = 1$ and $\mathcal{G} = \{2 \to 1\}$; otherwise, we would have $\mathrm{rank}(\Theta_1 - \Theta_0) = 1$. While i_1 and $\mathcal{G}$ are identifiable, preliminary computational evidence suggests that the entries of B_0, B_1, and H are not identifiable.*

Proof of Theorem 1. The necessity of d interventions is Prop. 5. Under Assumption 1 and 2, the sufficiency of d interventions follows from Theorem 2. Under Assumption 1, the sufficiency is the recovery of $\prec_{\mathcal{G}}$ in Prop. 3. □

4. Experimental Results

We adapt our proof of Theorem 3 into a method for causal disentanglement in the finite-sample setting. We modify our methods to (1) use matrices instead of vector spaces, (2) use scores based on singular values to test rank, and (3) choose a nonzero row based on norms. The adapted algorithms are in Appendix L. In this section, we investigate the performance of the method in a simulation study. There is a single hyperparameter $\gamma \in [0, 1]$, the percentage of spectral energy associated to the largest singular value, above which we consider a matrix to have rank one. We use $\gamma = 0.99$.
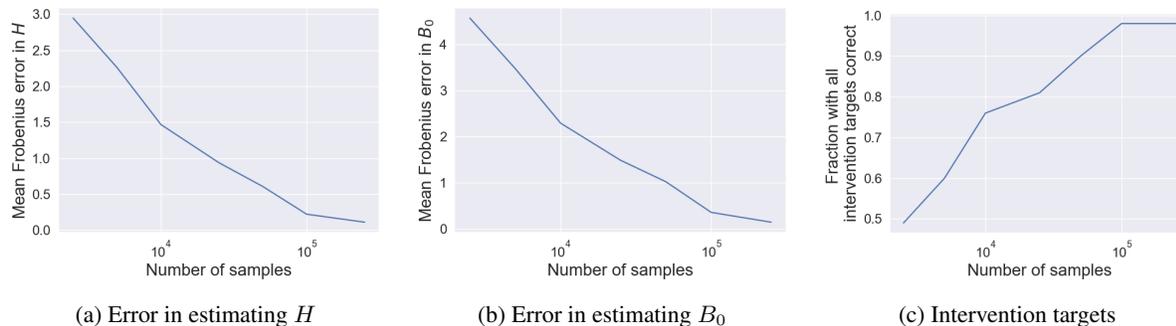
Figure 2: **The adapted version of Algorithm 3 is consistent for recovering H, B_0, and $\{i_k\}_{k=1}^K$ from noisy data.** At each sample size, we generate 500 random models. Note the logarithmic scale on the x-axis. In **(a)**, we plot the median of $\|\widehat{H} - H\|_2$, the error in Frobenius norm. In **(b)**, we plot the median of $\|\widehat{B}_0 - \widehat{B}\|_2$. In **(c)**, we plot the fraction of models where all intervention targets were correctly estimated.

4.1. Synthetic Data Generation

We generate 500 random models following Assumption 1 for $d = 5$ latent and $p = 10$ observed variables, as follows. We sample the graph $\mathcal{G}$ from an Erdős-Rényi random graph model with density 0.75. We sample the nonzero entries of A_0 independently from $\mathsf{Unif}(\pm[0.25, 1])$, and the nonzero entries of Ω_0 independently from $\mathsf{Unif}([2, 4])$. We sample uniformly among permutations to generate the intervention targets i_k. In context k, we have $A_k = A_0 - \mathbf{e}_{i_k} A_0^\top \mathbf{e}_{i_k}$; i.e., all entries in row i_k are 0. We change $(\Omega_0)_{i_k, i_k}$ into a new value $(\Omega_k)_{i_k, i_k}$, sampled from $\mathsf{Unif}([6, 8])$ to ensure a non-negligible change. Finally, the entries of H are sampled independently from $\mathsf{Unif}([-2, 2])$. We consider sample sizes n from 2500 to 250000 and use as input the sample precision matrices. All code for data generation and for our adapted versions of Algorithms 1, 2, and 3 (that is, Algorithms 6, 5 and 7) can be found at the link in Appendix M.

4.2. Synthetic Data Results

We show the results of applying our method in the finite-sample setting, presented in Fig. 2. We measure the error in estimating the parameters H and B_0 and the intervention targets $\{i_k\}_{k=1}^K$. Since the problem is only identifiable up to the partial order $\prec_{\mathcal{G}}$, we align our estimated $\widehat{H}$, $\widehat{B}_0$, and $\{\widehat{i}_k\}_{k=1}^K$ with H, B_0, and $\{i_k\}_{k=1}^K$ by picking $\sigma \in S(\mathcal{G})$ to maximize $\sum_{k=1}^K \mathbb{1}\{i_k = \sigma(\widehat{i}_k)\}$. For small d, this optimization can be solved by enumerating over $S(\mathcal{G})$. For large d, we use the integer linear program in Appendix M.1, implemented in $\texttt{gurobipy}$. As expected, the estimation errors for H and B_0 decrease as the sample size increases, while the fraction of models with all intervention targets correctly estimated increases.

4.3. Biological Data Results

We evaluate our method on a dataset from Ursu et al. (2022). This single-cell RNA sequencing (scRNA-seq) dataset consists of 90,000 cells from a lung cancer cell line, with 83 different nonsynonymous mutations of the KRAS oncogene overexpressed.

Semi-synthetic analysis. The authors divide the mutations into five categories based on the genes that they affect, and compute a score for the impact of each mutation. Taking the two highest impact mutations from each category gives $K = 10$ contexts. The wild type KRAS gene is taken as the observational context. We select $p = 100$ observational features to be the most variable genes from the proliferation regulation category in the Gene Ontology (Ashburner et al., 2000), which are significant modulators of cancer activity such as oncogenes and tumor suppressor genes. We compute the sample precision matrices $\widehat{\Theta}_0, \widehat{\Theta}_1, \ldots, \widehat{\Theta}_{10}$ and use them as input to Algorithm 7 with $\gamma = 0.99$.

Given estimates $\{\widehat{B}_k\}_{k=1}^K$, $\widehat{H}$ from our algorithm, we let $\tilde{B}_k = M_k \odot \widehat{B}_k$, where $(M_k)_{ij} = \mathbb{1}_{\{|(\widehat{B}_k)_{ij}|>0.04\}}$; i.e., we truncate the entries away from zero. We treat the resulting parameters as a new generative model. This tests our method in a more realistic setting, with parameters based on real data, while retaining the ability to measure performance. Since the entries of the matrices $\tilde{B}_k$ are smaller than for our synthetic data, we consider larger sample sizes $n \in \{1e6, 5e6, 1e7, 5e7, 1e8\}$. As seen in Fig. 6, Appendix M.2, we successfully recover the generative model.

Hypothetical Workflow. We illustrate a hypothetical workflow of our method on biological data. If we run our algorithm for all mutations ($K = 83$) on the $p = 83$ most variable genes, we obtain the graph in Fig. 3. We see that G12R, G12V, G13R, G13V, and G12I all perturb highly-connected latent nodes with several descendants. The G12 and G13 positions in the KRAS protein are key functional
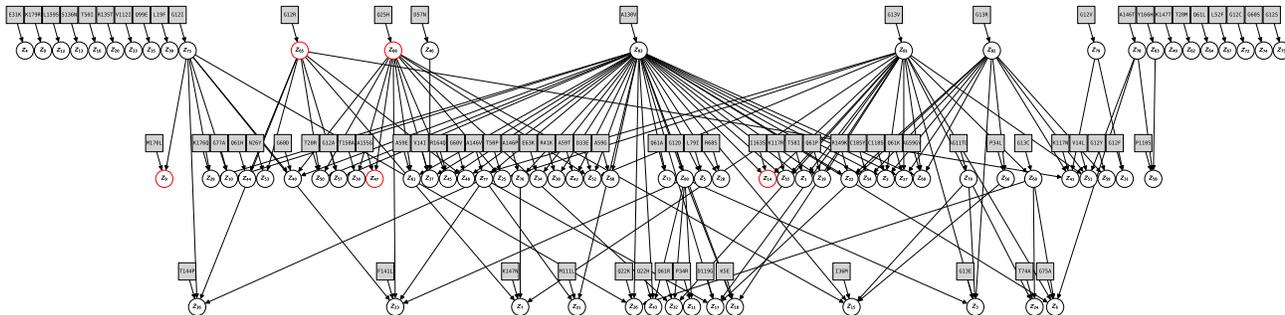
Figure 3: **The latent graph and intervention targets learned from scRNA-seq data.** Edges with weight of magnitude above 0.2 are shown. Boxes represent context indicators, corresponding to KRAS mutations, with edges to their respective targets. Red nodes are significantly associated with survival outcomes in the TCGA dataset.

residues whose mutations are known to be causal drivers of cancer (Huang et al., 2021). This indicates that the learned graph can highlight influential biological pathways which may be useful for prioritizing therapeutic development. The matrix $\widehat{H}$ from our algorithm gives a mapping from genes to latent variables that can be transferred across datasets and related to other observations. For example, we compute estimates of the latent variables for the 589 lung cancer patients in the Cancer Genome Atlas (Liu et al., 2018) and relate these variables to the patients' survival outcomes. We find that 5 latent variables, those targeted by the M170L, I163S, A55G, G12R, and Q25H mutations, are significantly associated with survival outcomes (at significance 0.1, after multiple testing correction). See Appendix M.2 for details. Note that the output of our method should be treated as exploratory; further theoretical and methodological development is required to better match complex real-world data.

5. Discussion

In this paper, we showed that a latent causal model is identifiable when data is available from an intervention on each latent variable. Conversely, we showed that, in the worst case, such data is necessary for identifiability of the latent representation. Our proof is constructive, consisting of an algorithm for recovering the latent representation, which can be adapted to the noisy setting. Our algorithm was developed for maximal clarity of our identifiability result, and leaves open several directions for future work.

Theory of latent interventional Markov equivalence. We established sufficient and (worst-case) necessary conditions for the *complete* identifiability of the parameters H and $\{B_k\}_{k=0}^K$. However, in many settings, it is expected that these parameters (and corresponding combinatorial structures) are only *partially* identifiable. Indeed, Proposition 5 suggests that the problem parameters may be partially recoverable. In future work, it would be interesting to develop a theory of identifiability for $K < p$ interventions.

Non-linear setting. Our results require that both the latent linear structural equation model and the mixing function are linear. We expect that the insights developed here may apply when one or both of these elements are non-linear. Notably, the contemporaneous work of Ahuja et al. (2022b) shows that, under certain conditions, the case of *polynomial mixing* can be reduced to the case of linear, which can be immediately applied to extend our result.

Statistical analysis of causal disentanglement. A next step beyond identifiability is to investigate the statistical properties of the setup. This includes lower bounds on the accuracy of recovering the parameters H and $\{B_k\}_{k=0}^K$, along with corresponding combinatorial structures such as $\mathcal{G}$ and the matching between k and i_k, and computationally efficient algorithms that match these lower bounds. Our identifiability result suggests that the differences of precision matrices may play a role. These differences appear in Varici et al. (2021), which uses techniques for directly estimating differences between precision matrices. Moreover, it would be interesting to develop a score-based approach, e.g., (penalized) maximum likelihood estimation. Our problem formulation suggests a natural parameterization for such an approach, which could be addressed using combinatorial optimization or gradient-based search.

Acknowledgements

References

Adams, J., Hansen, N., and Zhang, K. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34: 22822–22833, 2021.

Ahuja, K., Hartford, J., and Bengio, Y. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2021.

Ahuja, K., Hartford, J., and Bengio, Y. Weakly supervised representation learning with sparse perturbations. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=6ZI4iF_T7t.

Ahuja, K., Wang, Y., Mahajan, D., and Bengio, Y. Interventional causal representation learning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022b.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Brehmer, J., Haan, P. D., Lippe, P., and Cohen, T. Weakly supervised causal representation learning. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. URL https://openreview.net/forum?id=rUXOUBuUcg5.

Cai, R., Xie, F., Glymour, C., Hao, Z., and Zhang, K. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

Castelletti, F. and Peluso, S. Network structure learning under uncertain interventions. *Journal of the American Statistical Association*, pp. 1–12, 2022.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Davidson-Pilon, C. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

Drton, M., Sturmfels, B., and Sullivant, S. Algebraic factor analysis: tetrads, pentads and beyond. *Probability Theory and Related Fields*, 138:463–493, 2007.

Drton, M., Massam, H., and Olkin, I. Moments of minors of wishart matrices. 2008.

Eberhardt, F., Glymour, C., and Scheines, R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 178–184, 2005.

Greville, T. N. E. Note on the generalized inverse of a matrix product. *Siam Review*, 8(4):518–521, 1966.

Halpern, Y., Horng, S., and Sontag, D. Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299*, 2015.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Huang, L., Guo, Z., Wang, F., and Fu, L. Kras mutation: from undruggable to druggable in cancer. *Signal transduction and targeted Therapy*, 6(1):386, 2021.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Jaber, A., Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34: 18087–18101, 2021.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., et al. An integrated tcga

pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Hengel, A. v. d., Zhang, K., and Shi, J. Q. Weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. M. Identifiable deep generative models via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021.

Saeed, B., Belyaeva, A., Wang, Y., and Uhler, C. Anchored causal inference in the presence of measurement error. In *Conference on uncertainty in artificial intelligence*, pp. 619–628. PMLR, 2020.

Seabold, S. and Perktold, J. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., and Chickering, D. M. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7 (2), 2006.

Squires, C. and Uhler, C. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 2022.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

Squires, C., Yun, A., Nichani, E., Agrawal, R., and Uhler, C. Causal structure discovery between clusters of nodes induced by latent factors. In *Conference on Causal Learning and Reasoning*, pp. 669–687. PMLR, 2022.

Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Rätsch, G., and Lehmann, K.-V. Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36(Supplement_2):i919–i927, 2020.

Trefethen, L. N. and Bau III, D. *Numerical linear algebra*, volume 50. Siam, 1997.

Ursu, O., Neal, J. T., Shea, E., Thakore, P. I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M. M., et al. Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nature Biotechnology*, pp. 1–10, 2022.

Varici, B., Shanmugam, K., Sattigeri, P., and Tajer, A. Scalable intervention target estimation in linear models. *Advances in Neural Information Processing Systems*, 34: 1494–1505, 2021.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1990.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in Neural Information Processing Systems*, 33:14891–14902, 2020.

Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., and Zhang, K. Identification of linear non-Gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pp. 24370–24387. PMLR, 2022.

Yang, K., Katcoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

Contents of Appendix

A. Additional related work

Fig. 4 shows the two graphical conditions assumed in some prior works.

(a) Pure children   (b) No twins
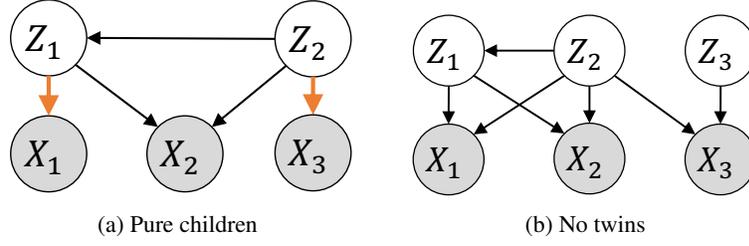
Figure 4: **Graphical conditions assumed in prior works.** In **(a)**, the orange edges link pure children (X_1 and X_3) to their parents (Z_1 and Z_3, respectively). In **(b)**, the no twins assumption is satisfied since the observed children of Z_1, Z_2, and Z_3 are, respectively, $\{X_1, X_2\}$, $\{X_1, X_2, X_3\}$, and $\{X_3\}$, and these three sets are distinct.

B. Non-generic soft interventions

We discuss the genericity condition in Assumption 1(b). We show that for soft interventions in which this genericity condition fails to hold, identifiability of the causal disentanglement problem as in Theorem 1 may fail. The following matrices satisfy all of Assumption 1, except for the genericity condition in Assumption 1(b), since $B_1^\top \mathbf{e}_1 = 2B_0^\top \mathbf{e}_1$:

$$B_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Consider the alternative matrices

$$\widehat{B}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \widehat{B}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \widehat{B}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \qquad \widehat{G} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

These do not differ from the original tuple of matrices via a permutation. However, one can check that they are a valid solution, since

$$\Theta_0 = \widehat{\Theta}_0 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \Theta_1 = \widehat{\Theta}_1 = \begin{bmatrix} 4 & 4 \\ 4 & 5 \end{bmatrix}, \qquad \Theta_1 = \widehat{\Theta}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}.$$

C. Pseudoinverse of a covariance matrix

Proposition 6. *Let $X = GZ$ for $G \in \mathbb{R}^{p \times d}$ with full column rank. Assume $Cov(Z)$ is invertible and let $K := Cov(Z)^{-1}$. Then $Cov(X)^\dagger = H^\top K H$, where $H := G^\dagger$.*

Proof. The covariance matrices $Cov(X)$ and $Cov(Z)$ are related via $Cov(X) = G \cdot Cov(Z) \cdot G^\top$. The property $(UV)^\dagger = V^\dagger U^\dagger$ holds whenever U has full column rank and V has full row rank (Greville, 1966). The matrix G has full column rank, $Cov(Z)$ has full rank, and $G^\top$ has full row rank. Hence $Cov(X)^\dagger = (G^\top)^\dagger Cov(Z)^\dagger G^\dagger = H^\top K H$. $\square$

D. The partial order RQ decomposition

Recall the partial order RQ decomposition from Definition 1. We present Algorithm 4 to find the partial order RQ decomposition of a matrix. In Line 4, the normalize operator is $\text{normalize}(\mathbf{v}) := \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$. We let $\mathbf{0}_{d \times p}$ denote the $d \times p$ matrix of zeros and $Q_{j \succeq i}$ denote the submatrix of Q on rows j with $j \succeq i$. We say that a partial order $\prec$ is consistent with the total order $1, 2, \ldots, d$ if $i \prec j$ implies $i < j$. Any partial order can be put in this form by relabelling.

Proposition 7. *Let $H \in \mathbb{R}^{d \times p}$ be full rank and fix a partial order $\prec$ over $[d]$. Then there exists a unique partial order RQ decomposition of H. If $\prec$ is consistent with the total order $1, 2, \ldots, d$, the decomposition is returned by Algorithm 4.*

Proof. The matrix $Q_{j \succeq i}$ has fewer rows than columns and $\mathbf{h}_i \in \langle \mathbf{q}_j : j \succeq i \rangle$, by its construction in Algorithm 4. Hence the vector $\mathbf{r}$ of non-zero entries in the i-th row of R is the unique solution to $Q_{j \succeq i}^\top \mathbf{r} = \mathbf{h}_i$. By construction, we see that $H = RQ$ and, moreover, $R_{ij} = 0$ if $j \not\succeq i$, and $\mathbf{q}_i$ is orthogonal to $\mathbf{q}_j$ for $i \prec j$. Furthermore, the entry R_{ii} is positive, since $\mathbf{q}_i$ is the (normalized) projection of $\mathbf{h}_i$ onto W_i. $\qquad\square$

Algorithm 4 Partial Order RQ Decomposition

1: **Input:** Matrix $H \in \mathbb{R}^{d \times p}$, partial order $\prec$ over $[d]$ consistent with the total order $1, 2, \ldots, d$
2: **Output:** A partial order RQ decomposition R, Q
3: Let $R = \mathbf{0}_{d \times d}$, $Q = \mathbf{0}_{d \times p}$
4: **for** $i = 1, \ldots, d$ **do**
5:      Let $\mathbf{h}_i$ be the i-th row of H
6:      Let $W_i = \langle \mathbf{q}_j : j \succ i \rangle$
7:      Let $\mathbf{q}_i = \text{normalize}(\text{proj}_{W_i^\perp} \mathbf{h}_i)$ be the i-th row of Q
8:      Let $\mathbf{r} = (Q_{j \succeq i}^\top)^\dagger \mathbf{h}_i$
9:      For $i \preceq j$, let $R_{ij} = r_j$
10: **end for**
11: **return** R, Q

E. Further preliminaries for identifiability and reduction

We prove Fact 1 and discuss the reduction described in Section 3.1.

Proof of Fact 1. We have $B^\top B = B^\top \left(\sum_{i=1}^d \mathbf{e}_i \mathbf{e}_i^\top \right) B = \sum_{i=1}^d B^\top (\mathbf{e}_i \mathbf{e}_i^\top) B = \sum_{i=1}^d (B^\top \mathbf{e}_i)^{\otimes 2}.$ $\qquad\square$

Proposition 8. *Consider the setup in Assumption 1. For $k, \ell \in [K]$, we have*

$$B_k^\top B_k - B_\ell^\top B_\ell = (B_k^\top \mathbf{e}_{i_k})^{\otimes 2} - (B_0^\top \mathbf{e}_{i_k})^{\otimes 2} - (B_\ell^\top \mathbf{e}_{i_\ell})^{\otimes 2} + (B_0^\top \mathbf{e}_{i_\ell})^{\otimes 2}, \tag{4}$$

and thus

$$\Theta_k - \Theta_\ell = (HB_k^\top \mathbf{e}_{i_k})^{\otimes 2} - (HB_0^\top \mathbf{e}_{i_k})^{\otimes 2} - (HB_\ell^\top \mathbf{e}_{i_\ell})^{\otimes 2} + (HB_0^\top \mathbf{e}_{i_\ell})^{\otimes 2}.$$

Proof. The proof follows the same steps as Prop. 1, together with the fact that

$$B_k^\top B_k - B_\ell^\top B_\ell = (B_k^\top B_k - B_0^\top B_0) - (B_\ell^\top B_\ell - B_0^\top B_0). \qquad\square$$

F. Reduction

In this section, we show that a more general causal disentanglement problem can be simplified to one that satisfies our assumptions. We consider an unknown observational context, and multiple contexts with the same target. We focus here on the case of perfect interventions.

F.1. Reducing to one intervention per node

We identify which contexts correspond to interventions on the same node. Thus, we can reduce to the case of one intervention per node by removing any redundant contexts. We do not use knowledge of which context is the observational context here.

Proposition 9. *Consider the setup in Assumptions 1 and 2. Assume generic parameters for B_0, λ_k, and λ_ℓ. For $k, \ell \in [K]$, we have $r_{k,\ell} = 1$ if and only if $i_k = i_\ell$, where $r_{k,\ell} := \text{rank}(\Theta_k - \Theta_\ell)$.*

Proof. Since H is full rank, we have $r_{k,\ell} = \text{rank}(B_k^\top B_k - B_\ell^\top B_\ell)$. Thus, we consider $\text{rank}(B_k^\top B_k - B_\ell^\top B_\ell)$. Suppose $i_k = i_\ell = i$. We have

$$B_k^\top B_k - B_\ell^\top B_\ell = \left(B_k^\top \mathbf{e}_i \right)^{\otimes 2} - \left(B_\ell^\top \mathbf{e}_i \right)^{\otimes 2},$$

by Equation (4) in Prop. 8. Both $B_k^\top \mathbf{e}_i$ and $B_\ell^\top \mathbf{e}_i$ have a single nonzero entry at the i-th coordinate, by Assumption 2. Thus $r_{k,\ell} = 1$.

Suppose $i_k \neq i_\ell$ and assume without loss of generality that $i_k < i_\ell$. Given a matrix M, let M_U denote the submatrix of M with rows and columns indexed by the elements of the set U. We have

$$\left(B_k^\top B_k - B_\ell^\top B_\ell\right)_{\{i_k, i_\ell\}} = \begin{bmatrix} \lambda_k^2 - (B_0)_{i_k, i_k}^2 & -(B_0)_{i_k, i_k}(B_0)_{i_k, i_\ell} \\ -(B_0)_{i_k, i_k}(B_0)_{i_k, i_\ell} & -\lambda_\ell^2 + (B_0)_{i_\ell, i_\ell}^2 - (B_0)_{i_k, i_\ell}^2 \end{bmatrix},$$

by Equation (4) in Prop. 8 and Assumption 2. For generic parameters, this submatrix has rank two, so the full matrix has rank at least two; i.e., $r_{k,\ell} \geq 2$. $\qquad\square$

F.2. Reducing to a known observational context

The previous section explains how to reduce to the case with one intervention per latent node. We may also reduce to the case with only one observational context: if more than one context is the observational context, they will all have the same inverse covariance matrix, so we may select only one of these contexts to serve as the observational context $k = 0$. Next we show that, with one intervention per node, and one observational context, we can identify the observational context. We show that the observational context has the "sparsest" changes from the other contexts. We formalize this intuition with the following definition.

Definition 2. *The* deviation score *of context k is*

$$r_k := \sum_{\ell \in [K] \setminus \{k\}} r_{k,\ell},$$

where $r_{k,\ell} := \mathrm{rank}(\Theta_k - \Theta_\ell)$ for all $k, \ell \in [K]$.

Proposition 10. *Consider the setup in Assumption 1. Then $k^* \in \{0\} \cup [K]$ is an observational context if and only if $k^* = \arg\min_{k \in \{0\} \cup [K]} r_k$.*

Proof. Let $\mathrm{source}(\mathcal{G})$ denote the set of source nodes in $\mathcal{G}$. By Prop. 1, $r_{0,\ell} = 1 + \mathbb{1}_{\mathrm{pa}(i_\ell) \neq \varnothing}$ for all $\ell \in [K]$. Thus, $r_0 = 2K - |\mathrm{source}(\mathcal{G})|$.

For $k \neq 0$, we have $r_{k,\ell} \geq 2 + \mathbb{1}_{\mathrm{pa}(i_\ell) \neq \varnothing}$ for all $\ell \in [K] \setminus \{k\}$. Thus, $2K$. Since $\mathcal{G}$ must have at least one source node, we see that $r_k > r_0$ for all $k \neq 0$. $\qquad\square$

G. Hypothesis testing a necessary condition for model membership

We define the null hypothesis

$$H_0 : \mathrm{rank}(\Theta_k - \Theta_0) \leq 2 \qquad \forall \, k \in [K]$$

Assumption 1(b) implies that H_0 holds, by Prop. 1. The null hypothesis H_0 is a necessary condition for membership of $(\Theta_0, \Theta_1, \ldots, \Theta_K)$ in the model defined by Assumption 1. However, H_0 is not a sufficient condition for model membership: we may have $\mathrm{rank}(\Theta_k - \Theta_0) \leq 2$ for all $k \in [K]$, despite some interventions not targeting single nodes. For example, if $\mathcal{G}$ is the empty graph, and all interventions have two targets, then H_0 holds. These cases may be ruled out with other conditions implied by model membership. We leave a membership test for our model to future work. Here, we focus on developing a test for H_0.

Prior work on testing latent variables models (Drton et al., 2007; Squires et al., 2022) use such rank constraints. To test whether a matrix $M \in \mathbb{R}^{p \times p}$ is rank k, one can test that all minors of size $k + 1$ vanish; i.e., the collection of hypotheses

$$H_{A,B} : t_{A,B} = 0 \qquad A, B \subseteq [p], |A| = |B| = k + 1$$

where $t_{A,B} := \det(M_{A,B})$.

For example, Squires et al. (2022) use this to test whether certain submatrices of a covariance matrix are rank one, as follows. Let $\widehat{M}$ be the sample covariance matrix computed from n samples. If the underlying distribution is multivariate Gaussian, it is well-known that $\widehat{M}$ follows a Wishart distribution. Now, for each pair of subsets A, B, compute the empirical minor
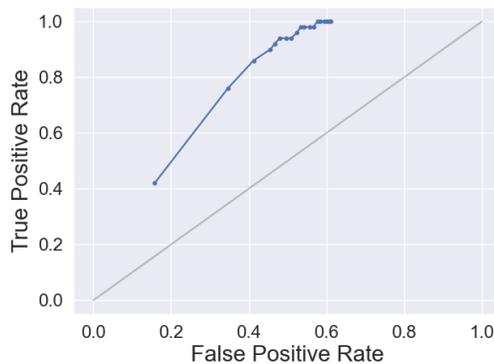
Figure 5: **Performance of our singular-value-based hypothesis test for H_0, the hypothesis that all precision matrix differences are of rank at most two.** The gray line indicates the performance of randomly guessing. Our test performs thresholding on the value $\rho_2(\widehat{\Theta}_k - \widehat{\Theta}_0)$. Here, we vary the threshold from 0.97 to 0.999.

$\widehat{t}_{A,B} := \det(\widehat{M}_{A,B})$. Then, compute an estimate $\widehat{\text{Var}}(t_{A,B})$. Such an estimate can be obtained by evaluating the expression for $\text{Var}(t_{A,B})$ in Drton et al. (2008), which characterizes the moments of minors for Wishart matrices. Given this estimate, compute the z-score $z_{A,B} = \widehat{t}_{A,B}/\sqrt{\widehat{\text{Var}}(t_{A,B})}$. By typical asymptotic theory, $z_{A,B} \to \mathcal{N}(0, 1)$ as $n \to \infty$, so we can use the z-score to compute an asymptotically correct p-value. Finally, the p-values for all pairs of subsets A, B can be aggregated into a single p-value using a multiple hypothesis testing procedure such as Bonferroni correction or Sidak adjustment.

In principle, a similar procedure can be performed to test our null hypothesis H_0. However, under a Gaussianity assumption, Θ_k and Θ_0 follow *inverse* Wishart distributions, rather than Wishart distribution. This would require expressions for the moments of minors for *inverse* Wishart matrices. We leave such a hypothesis test, which could give guarantees on false discovery rate control, to future work. Instead, we demonstrate the performance of a simple hypothesis test based on the singular values of the matrix $\widehat{\Theta}_k - \widehat{\Theta}_0$. Let

$$\rho_2(M) := \left(\sigma_1^2(M) + \sigma_2^2(M)\right) / \left(\sum_{i=1}^p \sigma_i^2(M)\right)$$

If $\text{rank}(M) \leq 2$, then $\rho_2(M) = 1$, otherwise, $\rho_2(M) < 1$. Thus, we may test H_0 by checking where $\rho_2(\widehat{\Theta}_k - \widehat{\Theta}) > \tau$ for some threshold τ near 1.

We demonstrate the performance of this procedure for testing model membership. We generate 500 random models following Assumption 1, using the same hyperparameters as in Section 4. These models satisfy H_0. We also generate 500 random models where the interventions target two nodes instead of one. For each k, we pick intervention targets $I_k \subset [d]$ with $|I_k| = 2$, uniformly at random among all subsets of size two. We hold all other hyperparameters of the simulation fixed. We consider only $n = 2500$ samples, the smallest sample size used in Section 4, and vary the threshold τ from 0.97 to 0.999, linearly spaced over 20 values. The singular value based test is able to determine model membership at a rate well above random guessing, see Fig. 5.

H. Computational complexity

Algorithm 3 takes as input K precision matrices in $\mathbb{R}^{p \times p}$. If these are each computed from at most n samples, then the total cost is $\mathcal{O}(Kp^3 + Knp^2)$. Algorithm 2 runs for K rounds and computes at most $2K$ projections per round. Each projection costs $\mathcal{O}(p^3)$, so the cost of this step is $\mathcal{O}(K^2p^3)$. In the remainder of Algorithm 3, we perform $\mathcal{O}(K)$ matrix multiplications and Cholesky decompositions. Each matrix multiplication costs $\mathcal{O}(p^2)$ and each Cholesky decomposition costs $\mathcal{O}(p^3)$. The other operations (e.g. selecting rows) in Algorithm 3 are negligible. Therefore, the overall runtime of Algorithm 3 is dominated by Algorithm 2, with a total cost of $\mathcal{O}(K^2p^3)$.

I. Proofs

I.1. Proof of Prop. 5

Proof. For (i), it suffices to find $\widehat{H}$ and $\{\widehat{B}_k\}_{k=1}^K$ such that $\widehat{B}_k \widehat{H} = B_k H$ for all $k \in [K]$, and such that there is no $\sigma \in S(\mathcal{G})$ satisfying $\widehat{B}_k = P_\sigma B_k P_\sigma^\top$, $H = P_\sigma H$, by Equation (2). Suppose $i_k \neq 1$ for any k. Let

$$\widehat{B}_k = \begin{bmatrix} \rule[0.5ex]{1em}{0.4pt}\;\; \mathbf{e}_1 \;\;\rule[0.5ex]{1em}{0.4pt} \\ (B_k)_{2:d,1:d} \end{bmatrix}, \qquad \widehat{H} = \begin{bmatrix} (B_0)_1^\top H \\ H_{2:d,1:p} \end{bmatrix}.$$

Then, for all k, we have $\widehat{B}_k \widehat{H} = B_k H$. Suppose that Z_1 has at least one parent, i.e., $(B_0)_{1j} \neq 0$ for at least one $j > 1$. Then the first row of $\widehat{H}$ is a nonzero combination of at least two rows of H. Hence it is not equal to a single row of H, since H is full rank. Thus, $\widehat{H}$ is not equal to H up to any permutation of rows.

For (ii), observe that for the stated example, the partial order $\prec$ given by $\widehat{B}_0$ differs in general from the partial order $\prec_{\mathcal{G}}$, since Z_1 has no predecessors in $\prec$. $\qquad\square$

J. Non-identifiability for imperfect interventions

J.1. Parameter non-identifiability

In this section, we show that Assumption 2 is necessary to identify H. Let

$$B_0 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} \\ 0 & (B_0)_{22} \end{bmatrix}, \qquad B_1 = \begin{bmatrix} (B_1)_{11} & (B_1)_{12} \\ 0 & (B_0)_{22} \end{bmatrix},$$

$$B_2 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} \\ 0 & (B_2)_{22} \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & H_{12} \\ 0 & 1 \end{bmatrix}.$$

Then, for any value $\widehat{H}_{12} \in \mathbb{R}$, we have $B_k H = \widehat{B}_k \widehat{H}$ for all k, where

$$\widehat{B}_0 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11} H_{12} - (B_0)_{11} \widehat{H}_{12} \\ 0 & (B_0)_{22} \end{bmatrix}, \qquad \widehat{B}_1 = \begin{bmatrix} (B_1)_{11} & (B_1)_{12} + (B_1)_{11} H_{12} - (B_1)_{11} \widehat{H}_{12} \\ 0 & (B_0)_{22} \end{bmatrix},$$

$$\widehat{B}_2 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11} H_{12} - (B_0)_{11} \widehat{H}_{12} \\ 0 & (B_2)_{22} \end{bmatrix}, \qquad \widehat{H} = \begin{bmatrix} 1 & \widehat{H}_{12} \\ 0 & 1 \end{bmatrix}.$$

J.2. Graph non-identifiability

Suppose that $\mathcal{G}$ has edges $2 \to 1$, $3 \to 2$, and $3 \to 1$. Then the weight matrices have the form

$$B_0 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_0)_{22} & (B_0)_{23} \\ 0 & 0 & (B_0)_{33} \end{bmatrix}, \qquad B_1 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_1)_{22} & (B_1)_{23} \\ 0 & 0 & (B_0)_{33} \end{bmatrix},$$

$$B_2 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_2)_{22} & (B_2)_{23} \\ 0 & 0 & (B_0)_{33} \end{bmatrix}, \qquad B_3 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} & (B_0)_{13} \\ 0 & (B_0)_{22} & (B_0)_{23} \\ 0 & 0 & (B_3)_{33} \end{bmatrix},$$

$$H = \begin{bmatrix} 1 & H_{12} & H_{13} \\ 0 & 1 & H_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $\widehat{\mathcal{G}}$ be the DAG with edges $2 \to 1$ and $3 \to 2$. We show that there exist $\widehat{H}$ and matrices $\widehat{B}_k$, following the support of $\widehat{\mathcal{G}}$, such that $B_k H = \widehat{B}_k \widehat{H}$ for all k. Note that $\mathcal{G}$ and $\widehat{\mathcal{G}}$ have the same transitive closure. Let $\widehat{H}_{12} \in \mathbb{R}$ and let $\widehat{H}_{13}, \widehat{H}_{23}$ be a solution to the system of equations:

$$\begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11}(H_{12} - \widehat{H}_{12}) \\ (B_1)_{11} & (B_1)_{12} + (B_1)_{11}(H_{12} - \widehat{H}_{12}) \end{bmatrix} \begin{bmatrix} \widehat{H}_{13} \\ \widehat{H}_{23} \end{bmatrix} = \begin{bmatrix} (B_0)_{11} H_{13} + (B_0)_{12} H_{23} + (B_0)_{13} \\ (B_1)_{11} H_{13} + (B_1)_{12} H_{23} + (B_1)_{13} \end{bmatrix}$$

This system generically has a solution, since for generic parameters the matrix on the left hand side is rank two. Then a solution, with all matrices $\widehat{B}_k$ have vanishing entry $(1, 3)$, is as follows.

$$\widehat{B}_0 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11}(H_{12} - \widehat{H}_{12}) & 0 \\ 0 & (B_0)_{22} & (B_0)_{23} + (B_0)_{22}(H_{23} - \widehat{H}_{23}) \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

$$\widehat{B}_1 = \begin{bmatrix} (B_1)_{11} & (B_1)_{12} + (B_1)_{11}(H_{12} - \widehat{H}_{12}) & 0 \\ 0 & (B_0)_{22} & (B_0)_{23} + (B_0)_{22}(H_{23} - \widehat{H}_{23}) \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

$$\widehat{B}_2 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11}(H_{12} - \widehat{H}_{12}) & 0 \\ 0 & (B_2)_{22} & (B_2)_{23} + (B_2)_{22}(H_{23} - \widehat{H}_{23}) \\ 0 & 0 & (B_0)_{33} \end{bmatrix}$$

$$\widehat{B}_3 = \begin{bmatrix} (B_0)_{11} & (B_0)_{12} + (B_0)_{11}(H_{12} - \widehat{H}_{12}) & 0 \\ 0 & (B_0)_{22} & (B_0)_{23} + (B_0)_{22}(H_{23} - \widehat{H}_{23}) \\ 0 & 0 & (B_3)_{33} \end{bmatrix}$$

K. Comparison to Liu et al. (2022)

We compare Theorem 2 to Liu et al. (2022). We restate their main result for convenience and notation.

Theorem (Theorem 4.1 of Liu et al. (2022)).   *Let $Z = A_k Z + \varepsilon_k$ and $X = g(Z) + \varepsilon_x$. Let $\boldsymbol{\eta}_k$ be the sufficient statistic for the distribution of Z in environment k. That is, $\boldsymbol{\eta}_k = \mathrm{vec}(\tilde{\Theta}_k)$, where $\tilde{\Theta}_k$ denotes the inverse covariance matrix of Z in the kth setting and vec denotes the vectorization of a matrix. We assume that vec ignores zeros and repetitions. Assume that*

(i) $\{x \in \mathcal{X} \mid \varphi_{\varepsilon_x}(x) = 0\}$ has measure zero, where φ_{ε_x} is the characteristic function for ε_x,

(ii) g is bijective, and

(iii) There exists $K + 1$ environments such that the following matrix is invertible:

$$L = \begin{bmatrix} | & | & & | \\ \boldsymbol{\eta}_1 - \boldsymbol{\eta}_0 & \boldsymbol{\eta}_2 - \boldsymbol{\eta}_0 & \dots & \boldsymbol{\eta}_K - \boldsymbol{\eta}_0 \\ | & | & & | \end{bmatrix}.$$

Then we can recover g up to permutation and scaling.

First, we show that (i) and (ii) hold in our setting. Our assumption that $X = GZ$ for G invertible guarantees (ii). Our assumption that X is a deterministic function of Z corresponds to taking $\varepsilon_x \sim \delta_0$, i.e., $\varepsilon_x = 0$ with probability one. The characteristic function is $\varphi_\varepsilon(t) = 1$, thus satisfying (i).

We now show that (iii) is only satisfied in our setting when the number of edges in the latent graph is at most d. The vector $\mathrm{vec}(\tilde{\Theta}_k - \tilde{\Theta}_0)$ is of length $d + |E|$, where $|E|$ is the number of edges in the graphical model defined by Θ_0. To be invertible, L must be a square matrix, and hence we require $K \geq d + |E|$. If $|E| > d$, then $K > 2d$, and we must have an intervention target i such that $i = i_k$ for at least three values of k. We have $\tilde{\Theta}_k = B_k^\top B_k$, and thus

$$\tilde{\Theta}_k - \tilde{\Theta}_1 = (\lambda_k \mathbf{e}_{i_k})^{\otimes 2} - (B_0^\top \mathbf{e}_{i_k})^{\otimes 2} - (\lambda_\ell \mathbf{e}_{i_1})^{\otimes 2} + (B_0^\top \mathbf{e}_{i_1})^{\otimes 2}$$

Given k_1, k_2, and k_3 such that $i_{k_1} = i_{k_2} = i_{k_3} = i$, we see that $\boldsymbol{\eta}_{k_1}, \boldsymbol{\eta}_{k_2}$ and $\boldsymbol{\eta}_{k_3}$ differ only at position (i, i). The space of vectors that differ in at most one entry is at most two-dimensional. Thus $\boldsymbol{\eta}_{k_1}, \boldsymbol{\eta}_{k_2}$, and $\boldsymbol{\eta}_{k_3}$ are not linearly independent, and L is not invertible.

L. Finite-sample algorithms

Matrix rank scoring. In Line 1 of Algorithm 1 and Line 2 of Algorithm 2, we check whether a subspace is rank one. In the finite-sample setting, we represent these subspaces by matrices and measure how close the matrices are to rank one.

Algorithm 5 IDENTIFYPARTIALORDERFINITESAMPLE

1: **Hyperparameters:** γ
2: **Input:** Precision matrices $(\Theta_0, \Theta_1, \ldots, \Theta_K)$
3: **Output:** Factor $\widehat{Q}$, partial order $\prec$
4: Let $\mathcal{I}_0 = \{\}$, $\widehat{Q} = \mathbf{0}_{d \times d}$
5: **for** $t = 1, \ldots K$ **do**
6:     Let $M_k = \mathrm{proj}_{\widehat{Q}_{t-1}^\perp}(\Theta_k - \Theta_0)$ for each $k \notin \mathcal{I}_{t-1}$
7:     Let $\rho_k = \sigma_1^2(M_k)/(\sum_{i=1}^p \sigma_i^2(M_k))$ for each $k \notin \mathcal{I}_{t-1}$
8:     Pick $k \in \arg\max_{k \notin \mathcal{I}_{t-1}} \rho_k$
9:     Let $\widehat{\mathbf{q}}_k, \mathcal{A} = $ IDENTIFYANCESTORSFINITESAMPLE$(\Theta_k, \Theta_0, \{\widehat{\mathbf{q}}_i\}_{i \in \mathcal{I}_{t-1}}; \gamma)$
10:     Add $a' \succ k$ for any $a' \succeq a, a \in \mathcal{A}$
11:     Let $\mathcal{I}_t = \mathcal{I}_{t-1} \cup \{k\}$, $\widehat{Q}_t = [\widehat{\mathbf{q}}_k; \widehat{Q}_{t-1}]$
12: **end for**
13: **return** $\widehat{Q}, \prec$

Algorithm 6 IDENTIFYANCESTORSFINITESAMPLE

1: **Hyperparameters:** γ
2: **Input:** $\Theta_k, \Theta_0, \{\widehat{\mathbf{q}}_i\}_{i \in \mathcal{I}}$
3: **Output:** Vector $\widehat{\mathbf{q}}_k$, ancestor set $\mathcal{A}$
4: Let $\mathcal{A} = \mathcal{I}$
5: **for** $i \in \mathcal{I}$ **do**
6:     Let $W_{\neg i} = [\widehat{\mathbf{q}}_i : j \in \mathcal{I} \setminus \{i\}]$
7:     Let $M_{\neg i} = \mathrm{proj}_{W_i^\perp}(\Theta_k - \Theta_0)$
8:     Let $\rho_k = \sigma_1^2(M_k)/(\sum_{i=1}^p \sigma_i^2(M_k))$ for each $k \notin \mathcal{I}_{-1}$
9:     If $\rho_k \geq \gamma$, let $\mathcal{A} = \mathcal{A} \setminus \{i\}$
10: **end for**
11: Let $W = [\widehat{\mathbf{q}}_a : a \in \mathcal{A}]$
12: Let $M = \mathrm{proj}_{W^\perp}(\Theta_k - \Theta_0)$
13: Let $\widehat{\mathbf{q}}_k$ be the (normalized) leading left singular vector of M
14: **return** $\widehat{\mathbf{q}}_k, \mathcal{A}$

We use the score $\rho(M) := \sigma_1^2(M)/\left(\sum_{i=1}^p \sigma_i^2(M)\right)$, where $\sigma_i(M)$ is the ith largest singular value of M; $\rho(M)$ can be interpreted as the percentage of *spectral energy* associated to the largest singular value of M. Using this score to choose the next element of the partial order does not require hyperparameters, see Line 5 of Algorithm 5. In contrast, using this score to prune the set of ancestors requires a hyperparameter to determine whether a matrix is close enough to rank one, see γ in Line 6 of Algorithm 6. Larger values of γ (e.g., 0.999) result in a more conservative algorithm and will output a denser latent graph, while smaller values of γ (e.g., 0.8) result in more aggressive pruning of the latent graph.

Picking $\mathbf{q}_k$. The matrix M in Line 6 of Algorithm 6 is not guaranteed to be rank one in the finite-sample case. We instead select the leading left singular vector of M.

Picking nonzero rows. In the finite-sample case, the matrix $\widehat{D}_k$ will not usually have only one nonzero row, see Line 3 of Algorithm 7. We estimate the intervention target i_k by picking the row of largest norm. Since we assume that i_k is distinct for distinct k, we maintain a set $\mathcal{N}$ of candidate intervention targets and do not allow replicates.

M. Code and Data

Our code can be found at

https://github.com/csquires/linear-causal-disentanglement-via-interventions.

Algorithm 7 ITERATIVEDIFFERENCEPROJECTIONFINITESAMPLE

1: **Hyperparameters:** γ
2: **Input:** Precision matrices $(\Theta_0, \Theta_1, \ldots, \Theta_K)$
3: **Output:** $\widehat{H}, (\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_K)$
4: Let $d = K$
5: Let $\widehat{Q}, \prec = $ IDENTIFYPARTIALORDERFINITESAMPLE$((\Theta_0, \Theta_1, \ldots, \Theta_K); \gamma)$
6: Let $\widehat{C}_k = $ CHOLESKY$((\widehat{Q}^\dagger)^\top \Theta_k \widehat{Q}^\dagger)$ for $k = 0, \ldots, K$
7: Let $\mathcal{N} = [p]$
8: Let $\widehat{R} = I_d$
9: **for** $k = 1, \ldots, K$ **do**
10:     Let $\widehat{D}_k = \widehat{C}_k - \widehat{C}_0$
11:     Pick $\widehat{i}_k \in \arg\max_{i \in \mathcal{N}} \|(\widehat{D}_k)_i\|_2$
12:     Let $\mathcal{N} = \mathcal{N} \setminus \{\widehat{i}_k\}$
13:     Let $\widehat{R}_{\widehat{i}_k} = (\widehat{D}_k)_{\widehat{i}_k} + (\widehat{C}_0)_{\widehat{i}_k}$
14: **end for**
15: Let $\widehat{H}' = \widehat{R}\widehat{Q}$
16: Let $\widehat{H} = \Lambda\widehat{H}'$, for Λ diagonal such that $\widehat{H}$ satisfies the conditions on H in Assumption 1(c)
17: Let $\widehat{B}_0 = $ CHOLESKY$((\widehat{H}^\dagger)^\top \Theta_0 \widehat{H}^\dagger)$
18: Let $\widehat{B}_k = \widehat{B}_0 + \mathbf{e}_{\widehat{i}_k} \left(|\widehat{\Lambda}_{\widehat{i}_k, \widehat{i}_k}| \mathbf{e}_{\widehat{i}_k} - \widehat{B}_0^\top \mathbf{e}_{\widehat{i}_k} \right)^\top$ for $k = 1, \ldots, K$
19: **return** $\widehat{H}, (\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_K)$

M.1. Optimizing over $S(\mathcal{G})$

Consider a partial order $\prec_\mathcal{G}$, a set of true intervention targets $i_1, \ldots, i_K$, and a set of estimated intervention targets $\widehat{i}_1, \ldots, \widehat{i}_K$. The integer linear program (5) computes the topological order π^* consistent with $\prec_\mathcal{G}$ that maximizes the number of agreements between i_k and $\widehat{i}_k$. The topological order π^* can be recovered by letting $\pi^*(i) = j$ for the unique j such that $A_{ij} = 1$. The first two lines of constraints ensure this uniqueness, and that $\pi^*(i) \neq \pi^*(i')$ for $i \neq i'$.

The final line of constraints ensures that π^* is consistent with $\prec_\mathcal{G}$. If π^* is not consistent with $\prec_\mathcal{G}$, then there exists i, i', j such that $i \prec_\mathcal{G} i'$, $A_{i'j} = 1$, and $A_{ij'} = 0$ for all $j' \leq j$, which violates the constraint $\sum_{j' \leq j} (A_{ij'} - A_{i'j'}) \geq 0$.

$$
\begin{aligned}
\max_{A_{ij} \in \{0,1\}^{d \times d}} \quad & \sum_{k=1}^{K} A_{i_k \widehat{i}_k} \\
\text{s.t.} \quad & \sum_{i=1}^{d} A_{ij} = 1 \quad \forall\, j \in [d] \\
& \sum_{j=1}^{d} A_{ij} = 1 \quad \forall\, i \in [d] \\
& \sum_{j' \leq j} (A_{ij'} - A_{i'j'}) \geq 0 \quad \forall\, i \prec_\mathcal{G} i', \forall\, j \in [d]
\end{aligned}
\tag{5}
$$

M.2. Additional information on real data

The scRNA-seq dataset of Ursu et al. (2022) is available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161824`. The TCGA dataset of Liu et al. (2018) is available at `https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LUAD.survival.tsv` and `https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-LUAD.htseq_fpkm.tsv.gz`.

Processing. We use EnrichR (Kuleshov et al., 2016) to pick the $p = 100$ and $p = 83$ most variable genes in the proliferation regulation gene set from the Gene Ontology. We use the values from the processed dataset from Ursu et al. (2022); the only additional processing removes cells which were assigned to synonymous mutations (i.e., those that do not

(a) Error in estimating H (b) Error in estimating B_0 (c) Intervention targets
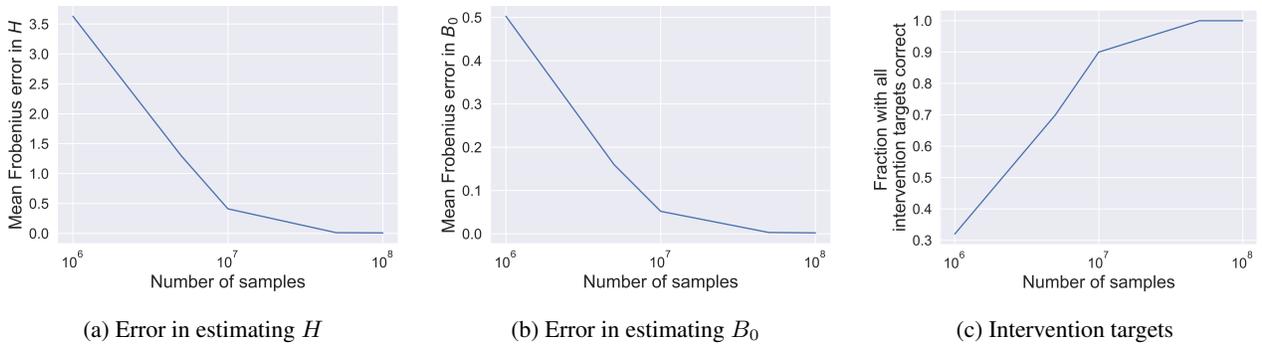
Figure 6: **(Semi-synthetic) The adapted version of Algorithm 3 is consistent for recovering H, B_0, and $\{i_k\}_{k=1}^K$ from semi-synthetic data.** At each sample size, we generate 50 datasets. Note the logarithmic scale on the x-axis. In **(a)**, we plot the mean of $\|\widehat{H} - H\|_2$, the error in Frobenius norm. In **(b)**, we plot the mean of $\|\widehat{B}_0 - B_0\|_2$. In **(c)**, we plot the fraction of models where all intervention targets were correctly estimated.

change any amino acids and hence do not have structural effects).

Semi-synthetic analysis. Our algorithm recovers the problem parameters for the semi-synthetic data, see Fig. 6.

Comparison to TCGA dataset. Our survival analysis is performed using the Cox proportional hazards model from the `lifelines` package (Davidson-Pilon, 2019). To correct for multiple hypothesis testing, we use the Benjamini-Hochberg procedure from the `statsmodels` package (Seabold & Perktold, 2010).