Learning Linear Causal Representations from General Environments: Identifiability and Intrinsic Ambiguity

Jikai Jin
Institute for Computational and Mathematical Engineering
Stanford University
Stanford, CA 94305
jkjin@stanford.edu

Vasilis Syrgkanis
Management Science and Engineering
Stanford University
Stanford, CA 94305
vsyrgk@stanford.edu

Abstract

We study causal representation learning, the task of recovering high-level latent variables and their causal relationships in the form of a causal graph from low-level observed data (such as text and images), assuming access to observations generated from multiple environments. Prior results on the identifiability of causal representations typically assume access to single-node interventions which is rather unrealistic in practice, since the latent variables are unknown in the first place. In this work, we consider the task of learning causal representation learning with data collected from *general environments*. We show that even when the causal model and the mixing function are both linear, there exists a *surrounded-node ambiguity* (SNA) [46] which is basically unavoidable in our setting. On the other hand, in the same linear case, we show that identification up to SNA is possible under mild conditions, and propose an algorithm, `LiNGCReL` which provably achieves such identifiability guarantee. We conduct extensive experiments on synthetic data and demonstrate the effectiveness of `LiNGCReL` in the finite-sample regime.

1 Introduction

Artificial intelligence (AI) has achieved tremendous success in various domains in the past decade [4, 40, 6]. However, current approaches are largely based on learning the *statistical* structures and relationships in the data that we observe. As a result, it is not surprising that these approaches often capture spurious statistical dependencies between different features, resulting in poor performance in the presence of test distribution shift [30, 22] or adversarial attacks [3, 50].

In view of these pitfalls, a recent line of work has explored the problem of *causal representation learning* (CRL) [34], the task of learning the causal relationships between high-level latent variables underlying our low-level observations. Notably, it is widely believed in cognitive psychology that humans take a causal approach to distill information from the world and make decisions to achieve their goals [37, 12, 19]. As a result, there is reason to believe that learning causal representations has the potential to significantly improve the power of AI, especially on tasks where performance lags far behind human level [17].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Despite such promise, a crucial challenge in CRL is the *identifiability* of the data generating process; in other words, given the data that we observe, can we uniquely identify the underlying causal model. It has been shown that given observational data (*i.e.*, i.i.d. data generated from a single environment), the model is already non-identifiable in strictly simpler settings where the latent variables are known to be independent [25, 26], or where there is no mixing function and one directly observes the latent variables [39]. As a result, existing algorithms for CRL with observational data [52, 53, 11] typically require additional assumptions on the structure of the underlying causal graph. A natural question that arises is what types of data do we need to acquire to make identification possible in the general case.

One line of works assumes access to counterfactual data [27, 48, 5], where some form of *weak supervision* is typically required. A common assumption here is that one observes data in *pairs*, where each pair of data is related via sharing part of the latent representation. However, such data is hard to acquire since it requires direct control on the latent representation.

Another line of works [1, 49, 7, 47] instead considers an interventional setting, where the learner observes data generated from multiple different environments. This is arguably a much more realistic setup and reflects common practices in robotics [24] and genomics [28, 43] applications. However, a vast majority of identifiability guarantees assume that each environment corresponds to *single-node*, *hard* interventions, which is defined as interventions that isolate a single latent variable from its causal parents. Again, this is quite a restrictive assumption because of two reasons. *First*, since the latent variables are unknown and need to be learned from data, it is unclear how to perform interventions that only affect one variable. *Second*, even if one can perform single-node interventions, it may not be feasible to artificially remove causal effects in the data generating processes. This issue is ubiquitous in real-world applications as pointed out in Campbell [8], Eberhardt [14], Eronen [15]. Motivated by these challenges, we make the following contributions in this paper:

- Assuming access to data collected from multiple environments, but not necessarily from single-node, hard interventions, we identify an intrinsic surrounded-node ambiguity (SNA) in learning the underlying causal representations. We show in Theorem 3 that SNA is unavoidable even if (1) both the mixing function and the causal model are known to be linear and (2) one has access to single-node, soft interventions. This highlights a remarkable difference with existing literature which shows that perfect identification can be achieved with hard interventions.
- When the causal model and the mixing function are both linear, we prove in Theorem 1 that identification up to SNA is achievable with $\mathcal{O}(d)$ diverse environments (Assumption 4), where d is the size of the latent causal graph. To the best of our knowledge, this is the first identification guarantee that applies to fully general environments and makes no assumption on their relationship or similarity. Interestingly, we also show in Theorem 2 that one would require $\Omega(d^2)$ single-node soft interventions to achieve the same identification guarantee, indicating the benefit of learning from diverse environments.
- We propose an algorithm, `LiNGCReL`, in Section 5 that provably recovers the ground-truth model up to SNA (Theorem 4) in the setting of Theorem 1 when perfect information of the observation distributions is available. To demonstrate the effectiveness of `LiNGCReL` in finite-sample regime, we conduct extensive experiments on synthetic data, and our results reported in Section 6 show that `LiNGCReL` is capable of recovering the true causal model up to SNA with high accuracy.

Due to space limit, proofs of all our statements and additional theoretical results are given in the appendix.

2 Preliminaries

We consider the standard setup of CRL from multiple environments $E \in \mathfrak{E}$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the ground-truth causal graph which is directed and acylic (DAG), where $\mathcal{V} = [d]$ and $\mathcal{E}$ describes the causal relationship between different nodes. Each node corresponds to a latent variable $\boldsymbol{z}_i \in \mathbb{R}$.

For any node i, we let $\mathrm{pa}_{\mathcal{G}}(i)$, $\mathrm{ch}_{\mathcal{G}}(i)$, $\mathrm{ans}_{\mathcal{G}}(i)$ and $\mathrm{nd}_{\mathcal{G}}(i)$ to be the set of all parents, children, ancestors and non-descendants of i in $\mathcal{G}$ respectively. We also define $\overline{\mathrm{pa}}_{\mathcal{G}}(i) = \mathrm{pa}_{\mathcal{G}}(i) \cup \{i\}$ and similarly for $\overline{\mathrm{ch}}_{\mathcal{G}}(i), \overline{\mathrm{ans}}_{\mathcal{G}}(i)$ and $\overline{\mathrm{nd}}_{\mathcal{G}}(i)$. Assuming that all probability distributions have continuous

densities, the joint density of the latent variables $\boldsymbol{z}$ can then be written as

$$p_E(\boldsymbol{z}) = \prod_{i=1}^{d} p_i^E \left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)} \right). \tag{1}$$

where p_i^E is the (unknown) latent generating distribution from environment E at node i. Here for a given vector $\boldsymbol{v}$, we write $\boldsymbol{v}_i = \boldsymbol{e}_i^\top \boldsymbol{v}$, and let $\boldsymbol{v}_S = (\boldsymbol{v}_i : i \in S) \in \mathbb{R}^{|S|}$.

The causal graph model with density given by (1) necessarily enjoys the following property:

Definition 1 (Causal Markov Condition)**.** *For any node i, conditioned on $\boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}$, $\boldsymbol{z}_i$ is independent of $\boldsymbol{z}_{\mathrm{nd}_{\mathcal{G}}(i)}$. As a consequence, for any node $i, j \in [d]$ and $S \subseteq [d]$, if S d-separates i from j (cf. Definition 7), then $\boldsymbol{z}_i \perp\!\!\!\perp \boldsymbol{z}_j \mid \boldsymbol{z}_S$.*

The latent variables $\boldsymbol{z}$ are unknown to the learner. Instead, the learner has access to observations $\boldsymbol{x} \in \mathbb{R}^n$ ($n \geqslant d$) from all environments $E \in \mathfrak{E}$ that are related to the latent $\boldsymbol{z}$ via an injective mixing function $\boldsymbol{g}$:

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z}). \tag{2}$$

The main assumption here that the mixing function is the same across all environments:

Assumption 1. *All environments $E \in \mathfrak{E}$ share the same diffeomorphic mixing function $\boldsymbol{g} : \mathbb{R}^d \mapsto \mathbb{R}^n$.*

In CRL, the goal of the learner is to 1) recover the inverse of the mixing function $\boldsymbol{h} = \boldsymbol{g}^{-1}$ (often called the *unmixing* function) which allows recovering the latent variables given any observations, and, 2) recover the underlying causal graph $\mathcal{G}$. In the remaining part of this paper, we refer to $(\boldsymbol{h}, \mathcal{G})$ as the causal model to be learned. Obviously, there would be some ambiguities in learning $(\boldsymbol{h}, \mathcal{G})$. For example, choosing a different permutation of the nodes in the causal graph would lead to a different model, and so does element-wise transformations on each component $\boldsymbol{h}_i$ of $\boldsymbol{h}$.

A line of recent works show that the ground-truth model can be identified up to these ambiguities in various settings, assuming access to single-node hard interventions [36, 49, 47]. On the other hand, some weaker notions of identifiability have also been proposed and studied in the literature [36, 46, 23] for single-node soft interventions. Here, we provide a generic definition of single-node soft interventions that we will rely on in this paper.

Definition 2. *We say that a collection of environments $\hat{\mathfrak{E}}$ is a set of (soft) interventions on a subset of latent variables $\{\boldsymbol{z}_j, j \in S\}$ if for any $i \in [d]$ and any $E_1, E_2 \in \hat{\mathfrak{E}}$, $E_1 \neq E_2$, we have $p_i^{E_1} = p_i^{E_2}$ if and only if $i \notin S$ (the notation p_i^E comes from (1)). Equivalently, we write $\mathcal{I}_{\boldsymbol{z}}^{\hat{\mathfrak{E}}} = S$.*

We note that soft interventions are very different from hard interventions, since they do not remove causal relationships between latent variables. The goal of this paper is to address the following question:

What is the best-achievable identification guarantee when hard interventions are not available, and what are the intrinsic ambiguities?

3 The surrounding set and a notion of identifiability

One may expect that identifiability with soft interventions is not much different from hard interventions, since soft interventions can approximate hard interventions with arbitrary accuracy. However, we will show that this is not the case. At a high level, hard intervention is more powerful than soft intervention because it is capable of isolating a latent variable from its direct cause while soft interventions is not, so soft interventions can sometimes fail to identify the true causal relationship from a mixture of causal effects.

To quantify what kind of ambiguities may arise, we can define the surrounding set for each node in a causal graph $\mathcal{G}$ as follows:

Definition 3. *(46, Definition 3) For two nodes $i, j \in [d]$ in $\mathcal{G}$, we say that j is surrounded by i, or $i \in \mathrm{sur}_{\mathcal{G}}(j)$ if $i \in \mathrm{pa}_{\mathcal{G}}(j)$, and $\mathrm{ch}_{\mathcal{G}}(j) \subseteq \mathrm{ch}_{\mathcal{G}}(i)$. Moreover, we define $\overline{\mathrm{sur}}_{\mathcal{G}}(j) = \mathrm{sur}_{\mathcal{G}}(j) \cup \{j\}$.*

Intuitively, if there exists some $i \in \mathrm{sur}_{\mathcal{G}}(j)$, then ambiguities may arise for the causal variable at node j, since any effect of j on any of its child k can also be interpreted as a mixture of the effect of i

and j. In Appendix E we discuss an example with three causal variables to further illustrate such ambiguities.

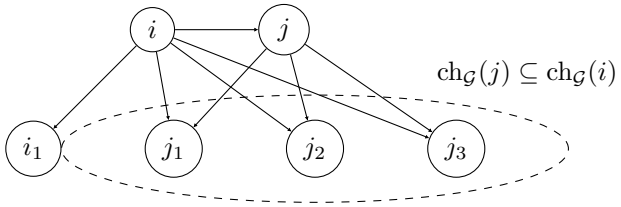


Figure 1: An illustration of Definition 3; here $i \in \mathrm{sur}_{\mathcal{G}}(j)$.

Definition 3 naturally induces the following relationship between causal models:

Definition 4. *Using the notations in Definition 10, we write $(\boldsymbol{h}, \mathcal{G}) \sim_{\mathrm{sur}} (\hat{\boldsymbol{h}}, \hat{\mathcal{G}})$ if there exists a permutations π on $[d]$, and a diffeomorphism $\psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ where the j-th component of ψ, denoted by $\psi_j(\boldsymbol{z})$, is a function of $\boldsymbol{z}_{\overline{\mathrm{sur}}_{\mathcal{G}}(j)}$ for $\forall j \in [d]$, such that the following holds:*

- *For any $i, j \in [d]$, $i \in \mathrm{pa}_{\mathcal{G}}(j)$ if and only if $\pi(i) \in \mathrm{pa}_{\hat{\mathcal{G}}}(\pi(j))$, and*
- *$\boldsymbol{P}_\pi \circ \hat{\boldsymbol{h}} = \psi \circ \boldsymbol{h}$, where $\boldsymbol{P}_\pi$ is a permutation matrix satisfying $(\boldsymbol{P}_\pi)_{ij} = 1$ if $j = \pi(i)$ and $(\boldsymbol{P}_\pi)_{ij} = 0$ otherwise.*

In other words, $\sim_{\mathrm{sur}}$ requires that the causal graph to be exactly the same up to some permutation of nodes, but allows each latent variable $\boldsymbol{v}_i$ to be a mixture of $\boldsymbol{z}_{\mathrm{sur}_{\mathcal{G}}(i)}$. Although not obvious from definition, one can actually check that $\sim_{\mathrm{sur}}$ defines an *equivalence relation* (see Lemma 11). Moreover, we will show in the following section that $\sim_{\mathrm{sur}}$ is in general the best that we can hope for in our problem setting.

4 Identifiability theory for linear CRL with general environments

In this section, we consider learning causal models from *general* environments. Specifically, we assume that the environments $E_k, k \in [K]$ share the same causal graph, but the dependencies between connected nodes (latent variables) are completely unknown, and, in contrast with existing literature on single-node interventions, we impose no similarity constraints on the environments. We begin our investigation of identifiability in this setting in the context of linear causal models with a linear mixing function.

4.1 Problem setup

Formally, we assume the following generative model in K distinct environments $\mathfrak{E} = \{E_k : k \in [K]\}$ with data generating process

$$\boldsymbol{z} = \boldsymbol{A}_k \boldsymbol{z} + \boldsymbol{\Omega}_k^{\frac{1}{2}} \epsilon, \quad \boldsymbol{x} = \boldsymbol{G}\boldsymbol{z} \quad k \in [K], \tag{3}$$

where the matrix $\boldsymbol{A}_k$ satisfies $(\boldsymbol{A}_k)_{ij} \neq 0$ if and only if $j \to i$ in $\mathcal{G}$. We refer to $(\boldsymbol{A}_k, \boldsymbol{\Omega}_k)$ as the weight matrices of latent variables $\boldsymbol{z}$ in the environment E_k. It is easy to see that Assumption 1 in our general setup translates into the following assumption:

Assumption 2. *The mixing matrix $\boldsymbol{G} \in \mathbb{R}^{n \times d}$ has full column rank. Equivalently, the unmixing matrix $\boldsymbol{H} = \boldsymbol{G}^\dagger$ has full row rank.*

Let $\boldsymbol{B}_k = \boldsymbol{\Omega}_k^{-\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{A}_k), k \in [K]$, then we have $\epsilon = \boldsymbol{B}_k \boldsymbol{z} = \boldsymbol{B}_k \boldsymbol{H} \boldsymbol{x}$. Since in the linear case, there is an easy to see one-to-one correspondence between the matrix $\boldsymbol{H}$ and the un-mixing function $\boldsymbol{x} \mapsto \boldsymbol{H}\boldsymbol{x}$, we abuse the notation and write $(\boldsymbol{H}, \mathcal{G})$ to represent the model instead of $(\boldsymbol{h}, \mathcal{G})$. Using $\boldsymbol{h}_i$ to denote the i-th row of $\boldsymbol{H}$, the following lemma translates Definition 4 the the linear setting:

Lemma 1. *According to Definition 4, $(\boldsymbol{H}, \mathcal{G}) \sim_{\mathrm{sur}} (\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$ if and only if there exists a permutation π on $[d]$, such that the following statements hold:*

1. *For all* $i, j \in [d]$, $i \in \mathrm{pa}_{\mathcal{G}}(j)$ *if and only if* $\pi(i) \in \mathrm{pa}_{\hat{\mathcal{G}}}(\pi(j))$, *and*

2. *For all* $i \in [d]$, $\hat{\boldsymbol{h}}_i \in \operatorname{span}\langle \boldsymbol{h}_j : \pi(j) \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)\rangle$.

We also need to make the following assumption on noise.

Assumption 3. *The noise vector* $\epsilon \in \mathbb{R}^d$ *has independent components, at most one component is Gaussian distributed, and any two components have different distribution.*

The non-gaussianity of the noise vectors is a typical assumption in causal discovery within linear models [9, 39] and is always assumed in the LinGAM setting [38]. The assumption that all components have a different distribution is not so standard, but is quite natural in real-world scenarios.

4.2 Identifiability guarantee

For each node $i \in [d]$ of $\mathcal{G}$, we use $\boldsymbol{w}_k(i)$ to be the *weight vector* of environment E_k at node i, *i.e.*, $\boldsymbol{w}_k(i) = \big((\boldsymbol{A}_k)_{ij} : j \in \mathrm{pa}_{\mathcal{G}}(i)\big) \in \mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|}$. In other words, the structural equation for node i in environment k is of the form:

$$z_i = w_k(i)^\top z_{\mathrm{pa}_{\mathcal{G}}(i)} + \sqrt{\omega_{k,i,i}}\epsilon_i \tag{4}$$

To obtain our identifiability result, the main assumption we need to make is the non-degeneracy of the weights at each node:

Assumption 4. *For each node* $i \in [d]$ *of* $\mathcal{G}$, *we have* $\operatorname{aff}(\boldsymbol{w}_k(i) : k \in [K]) = \mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|}$ *where* $\operatorname{aff}(\cdot)$ *denotes the affine hull. Equivalently, the weights* $\boldsymbol{w}_k(i), k = 1, 2, \cdots, K$ *do not lie in a* $\big(|\mathrm{pa}_{\mathcal{G}}(i)| - 1\big)$*-dimensional hyperplane of* $\mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|}$.

This assumption is quite mild since it only requires the weight vectors to be in general positions, and it holds with probability 1 if the weights at each node are sampled from continuous distributions. Moreover, as shown in Lemma 5, it is equivalent to the following assumption.

Assumption 5 (Node-level non-degeneracy)**.** *We say that the matrices* $\{\boldsymbol{B}_k\}_{k=1}^K$ *are node-level non-degenerate if for all node* $i \in [d]$, *we have* $\dim \operatorname{span}\langle (\boldsymbol{B}_k)_i : k \in [K]\rangle = |\mathrm{pa}_{\mathcal{G}}(i)| + 1$, *where* $(\boldsymbol{B}_k)_i$ *is the* i*-th row of* $\boldsymbol{B}_k$.

In the following, we state our main result in this section, which shows that $K = d$ non-degenerate environments suffices for the model to be identifiable up to $\sim_{\mathrm{sur}}$.

Theorem 1. *Suppose that* $K \geqslant d$ *and we have access to observations generated from the linear causal model* $(\boldsymbol{H}, \mathcal{G})$ *across multiple environments* $\mathfrak{E} = \{E_k : k \in [K]\}$ *with observation distributions* $\{\mathbb{P}_{\boldsymbol{x}}^E\}_{E\in\mathfrak{E}}$, *and the data generating processes are given by (3). Let* $(\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$ *be any candidate solution with the hypothetical data generating process*

$$\boldsymbol{v} = \hat{\boldsymbol{A}}_k \boldsymbol{v} + \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}} \hat{\epsilon}, \quad \boldsymbol{x} = \hat{\boldsymbol{H}}^\dagger \boldsymbol{v} \quad \textit{in the environment } E_k$$

where $\hat{\boldsymbol{H}}$ *has full row rank, such that*

(i) the observation distribution that this hypothetical model generates in E_k *is exactly* $\mathbb{P}_{\boldsymbol{x}}^{E_k}$;

(ii) all environments share the same causal graph: $\forall k \in [K]$ *and* $i, j \in [d]$, $(\boldsymbol{A}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\mathcal{G}}(i)$, $(\hat{\boldsymbol{A}}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\hat{\mathcal{G}}}(i)$ *and* $\boldsymbol{\Omega}_k, \hat{\boldsymbol{\Omega}}_k$ *are diagonal matrices with positive entries;*

(iii) $\{\boldsymbol{B}_k\}_{k=1}^K$ *and* $\Big\{\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{\Omega}}_k^{-\frac{1}{2}}(\boldsymbol{I} - \hat{\boldsymbol{A}}_k)\Big\}_{k=1}^K$ *are non-degenerate in the sense of Assumption 5;*

(iv) the noise variables ϵ *and* $\hat{\epsilon}$ *satisfy Assumption 3.*

Then we must have $(\boldsymbol{H}, \mathcal{G}) \sim_{\mathrm{sur}} (\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$.

The proof of Theorem 1 is given in Appendix H.1. In the next section, we will introduce an algorithm, `LiNGCReL`, that provably recovers the ground-truth up to $\sim_{\mathrm{sur}}$.

To the best of our knowledge, this is the first identifiability guarantee in the literature for CRL from general environments, even for the linear case. Our result is closely related but fundamentally different from Xie et al. [52, 53], Dong et al. [11] that consider the task of linear CRL using *observational data*. As discussed before, with observational data the causal graph can at best be identified up to Markov equivalence. As a result, one typically requires additional assumptions on the structure of the causal graph to obtain stronger guarantees. In contrast, we show that with data from multiple environments, exact recovery of the causal graph is possible without any structural assumptions.

Interestingly, while the fact that existing works focus on single-node interventions seem to suggest that learning from diverse environments is hard, it turns out that such diversity is actually helpful. Specifically, we show that in the worst case, $\Theta(d^2)$ interventions are required for identifying the ground-truth model under $\sim_{\text{sur}}$:

Theorem 2 (informal version of Theorem 6)**.** *There exists a causal graph $\mathcal{G}$ with $\Theta(d^2)$ edges, such that for any unmixing matrix $\boldsymbol{H} \in \mathbb{R}^{d\times n}$ with full row rank, any independent noise variables ϵ, and any $0 < s_i \leqslant \left|\text{pa}_{\mathcal{G}}(i)\right|, i \in [d]$, the ground-truth model $(\boldsymbol{H}, \mathcal{G})$ is non-identifiable up to $\sim_{\text{sur}}$ with s_i soft interventions for node i, unless the (ground-truth and intervened) weights of the causal model lie in a null set (w.r.t the Lebesgue measure).*

A formal version and the proof of Theorem 2 can be found in Appendix H.2. On the other hand, by having d single-node interventions per node, Assumption 5 can be satisfied as long as the weights are in general positions, so in this case we have $(\boldsymbol{H}, \mathcal{G}) \sim_{\text{sur}} (\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$ by Theorem 1. Therefore, Theorems 1 and 6 together imply that $\Theta(d^2)$ single-node interventions are necessary and sufficient for identification up to $\sim_{\text{sur}}$.

Given that Theorem 1 only guarantees identification up to $\sim_{\text{sur}}$ that is strictly weaker than full identification, one might naturally ask whether Theorem 1 can be further improved. Our last theorem in this section indicates that $\sim_{\text{sur}}$ is indeed a fundamental barrier that exists even when we access to *single node, soft interventions*.

Theorem 3 (Counterpart to Theorem 1, informal version of Theorem 9)**.** *For any linear causal model $(\boldsymbol{H}, \mathcal{G})$ and any set of environments $\mathfrak{E} = \{E_k : k \in [K]\}$ such that all conditions in Theorem 1 are satisfied, there must exists a candidate solution $(\hat{\boldsymbol{H}}, \mathcal{G})$ and a hypothetical data generating process that satisfy the same set of conditions, but*

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad \forall j \in \overline{\text{sur}}_{\mathcal{G}}(i).$$

Moreover, if we additionally assume that the environments are groups of single-node soft interventions, then we can guarantee the existence of $(\hat{\boldsymbol{H}}, \mathcal{G})$ and weight matrices which, besides the properties listed above, are also groups of single-node soft interventions.

5 `LinGCReL`: Algorithm for linear non-Gaussian causal representation learning

In this section, we introduce Linear Non-Gaussian Causal Representation Learning (`LiNGCReL`), an algorithm that provably recovers the underlying causal graph and latent variables up to $\sim_{\text{sur}}$ in the infinite-sample limit. At this point, it is instructive to recall the celebrated `LiNGAM` algorithm [38] for linear causal graph discovery. Different from their setting, we only observe some unknown linear mixture of the latent variables. Hence, running linear ICA as in `LiNGAM` only gives us $\boldsymbol{M}_k = \boldsymbol{B}_k \boldsymbol{H}$ rather than the weight matrix $\boldsymbol{B}_k$ itself.

The key idea in our approach is an effect cancellation scheme that allows us to determine the "remaining degree of freedom" (RDF) of any node (*a.k.a.* latent variable) given any subset of its ancestors. This scheme allows us to not only find a topological order of the nodes, but also figure out direct causes by tracking the changes of the RDF. In the following, we present the main steps of `LiNGCReL` in more details.

Suppose that we are given samples of observations $\boldsymbol{X}^{(k)} = \left\{\boldsymbol{x}_i^{(k)}\right\}_{i=1}^{N}, k \in [K]$ where $\boldsymbol{x}_i^{(k)}$ is the i-th sample from the k-th environment.

Step 1. Recover the matrices $\boldsymbol{M}_k = \boldsymbol{B}_k \boldsymbol{H}$ Since $\epsilon = \boldsymbol{B}_k \boldsymbol{z} = \boldsymbol{B}_k \boldsymbol{H} \boldsymbol{x}$ in the k-th environment, so we can use any identification algorithm for linear ICA to recover the matrix $\boldsymbol{M}_k$. Then we properly

rearrange the rows of $\boldsymbol{M}_k$ so that all $\boldsymbol{M}_k\boldsymbol{x}, k = 1, 2, \cdots, K$ correspond to the same permutation of noise variables. This step is quite standard and details can be found in Appendix B.1.

Step 2. CRL based on $\boldsymbol{M}_k$ Now we have obtained $\boldsymbol{M}_k = \boldsymbol{B}_k\boldsymbol{H}$, but the unmixing matrix $\boldsymbol{H}$ is still unknown. We propose Algorithm 3 to learn $\boldsymbol{H}$ and the causal graph $\mathcal{G}$. The main part of Algorithm 3 contains a loop that maintains a node set S which, we will show later, is ancestral, *i.e.,* $i \in S \Rightarrow \mathrm{ans}_{\mathcal{G}}(i) \subseteq S$. In each round the algorithm finds a new node $i \notin S$ such that $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S$, and a subroutine `Identify-Parents` (Algorithm 2) is used to find all parents of i. After that, we append i into S and continue until S contains all nodes in $\mathcal{G}$. Finally, the rows of the mixing matrix $\boldsymbol{H}$ is obtained by intersections of properly-chosen row spaces of $\boldsymbol{M}_k$.

Both Algorithm 2 and Algorithm 3 include a crucial step, which we call it *orthogonal projection*, as described in Algorithm 1. At a high level, it helps determine the minimal RDF for $\boldsymbol{z}_i$ after fixing the latent variables $\boldsymbol{z}_S$, and this exactly corresponds to the number of parents of $\boldsymbol{z}_i$ that are not in $\boldsymbol{z}_S$. We provide a simple example in Appendix E.2 to illustrate why this approach works.

The following result states that Algorithm 3 can recover the ground-truth causal model up to $\sim_{\mathrm{sur}}$:

Theorem 4. *Suppose that $\boldsymbol{M}_k, k \in [K]$ are perfectly identified in **Step 1**. Let $(\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$ be the solution returned by Algorithm 3, then we must have $(\boldsymbol{H}, \mathcal{G}) \sim_{\mathrm{sur}} (\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$.*

The full proof of Theorem 4 is given in Appendix H.3. It crucially relies on the following two propositions that reveal how Algorithm 3 and the subroutine Algorithm 2 work.

Algorithm 1 `Orthogonal-projections`

Input: Ordered set $S = \{s_1, s_2, \cdots, s_m\} \subseteq [d]$, index $i \notin S$, matrices $\boldsymbol{M}_k \in \mathbb{R}^{d\times n}$, $k \in [K]$
Output: Set of vectors $\{\boldsymbol{p}_k\}_{k=1}^K$
for $k \leftarrow 1$ to K **do**
  $\boldsymbol{W} \leftarrow \mathrm{span}\,\langle (\boldsymbol{M}_k)_s : s \in S\rangle$ ▷ $(\boldsymbol{M}_k)_s$ is the s-th row of $\boldsymbol{M}_k$
  $\boldsymbol{p}_k \leftarrow \mathrm{proj}_{\boldsymbol{W}^\perp}((\boldsymbol{M}_k)_i)$
end for

Proposition 1. *The following two propositions hold for Algorithm 3:*

- $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S \Leftrightarrow$ *the* `if` *condition in line 8 of Algorithm 3 is fulfilled;*
- *the set S maintained in Algorithm 3 is always an ancestral set, in the sense that $j \in S \Rightarrow \mathrm{ans}_{\mathcal{G}}(j) \subseteq S$.*

Proposition 2. *Given any ordered ancestral set S that contains $\mathrm{pa}_{\mathcal{G}}(i)$ for some $i \notin S$, Algorithm 2 returns a set $P_i \subseteq S$ that is exactly $\mathrm{pa}_{\mathcal{G}}(i)$.*

Algorithm 2 `Identify-Parents`

Input: An ordered set $S = \{s_1, s_2, \cdots, s_m\} \subseteq [d]$, a node $i \notin S$ and matrices $\boldsymbol{M}_k, k \in [K]$
Output: The parent set P_i of node i
$P_i \leftarrow \emptyset$
for $m' \leftarrow 0$ to m **do**
  $\{\boldsymbol{p}_k\}_{k=1}^K \leftarrow$ `Orthogonal-projections` $\left(\{s_j : j \leqslant m'\}, i, \{\boldsymbol{M}_k\}_{k\in[K]}\right)$
  $r_{m'} \leftarrow \dim \mathrm{span}\,\langle \boldsymbol{p}_k : k \in [K]\rangle$
  if $m' \geqslant 1$ and $r_{m'} = r_{m'-1} - 1$ **then**
    $P_i \leftarrow P_i \cup \{m'\}$
  end if
end for

6 Experiments

In this section, we present our experimental setup and results for `LiNGCReL`. Note that `LiNGCReL` as described in the previous section only works in the population regime. When the number of samples is limited, two main challenges in implementing `LiNGCReL` are to accurately compute the dimension

Algorithm 3 `Learn-Causal-Model`

Input: Matrices $\boldsymbol{M}_k, k \in [K]$
Output: The edge set $\mathcal{E}$ on the vertex set $[d]$ and the mixing matrix $\hat{\boldsymbol{H}}$
$S \leftarrow \emptyset$; ▷ S is an ordered set of nodes
$\mathcal{E} \leftarrow \emptyset$; ▷ $\mathcal{E}$ is the edge set
while $|S| < d$ **do**
  for $i \notin S$ **do**
    $\{\boldsymbol{p}_k\}_{k=1}^K \leftarrow$ `Orthogonal-projections` $\left(S, i, \{\boldsymbol{M}_k\}_{k\in[K]}\right)$
    if $\dim \operatorname{span}\langle \boldsymbol{q}_k : k \in [K]\rangle = 1$ **then**
      break ▷ Proposition 1 guarantees that such an i must exist
    end if
  end for
  $P_i \leftarrow$ `Identify-Parents`(S, i)
  $S \leftarrow S \cup \{i\}$
  $\mathcal{E} \leftarrow \mathcal{E} \cup \{(j, i) : j \in P_i\}$
end while
for $i = 1$ to d **do**
  $E_i \leftarrow \operatorname{span}\langle (\boldsymbol{M}_k)_i : k \in [K]\rangle$
end for
for $i = 1$ to d **do**
  $\hat{\boldsymbol{h}}_i \leftarrow$ any non-zero vector in $\left(\cap_{j:(i,j)\in\mathcal{E}} E_j\right) \cap E_i$
end for
$\hat{\boldsymbol{H}} \leftarrow \left[\hat{\boldsymbol{h}}_1^\top, \hat{\boldsymbol{h}}_2^\top, \cdots, \hat{\boldsymbol{h}}_d^\top\right]^\top$

of a subspace (line 6 of Algorithm 2 and line 8 of Algorithm 3), and to find a vector in the intersection of multiple subspaces (line 20, Algorithm 3). Due to space limit, the implementation details are described in Appendix B.2.

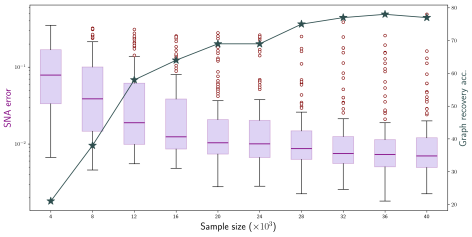
Experimental setup. We generate the independent noise variables from generalized Gaussian distributions $p_\beta(x) \propto \exp\left(-|x|^\beta\right)$ with parameters $\beta_k = 0.2k^2, k = 1, 2, \cdots, d$, multiplied by normalization constants to make their variances equal to 1. The ground-truth causal graph is generated by first fixing a total order of the vertices, say $1, 2, \cdots, d$, then add directed edges $i \to j (i < j)$ according to i.i.d. Bernoulli(p) distributions, where $p \in (0, 1)$. The non-zero entries of matrices $\boldsymbol{B}_k$ and $\boldsymbol{H}$ are all generated independently from Gaussian distributions. For simplicity, we focus on the case $n = d$ since recovery of the latent graphs only requires information from d components of $\boldsymbol{x}$.

Metrics of estimation error. Since CRL seeks to learn both the causal graphs and the latent variables, for each output of our algorithm we first check if it exactly recovers the ground-truth causal graph. Then, recall that the latent variables and the observations are related by $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$, given any output unmixing matrix $\hat{\boldsymbol{H}}$ from Algorithm 3, we define the relative estimation error Δ_i for $\boldsymbol{z}_i$ as the solution of the following optimization problem:

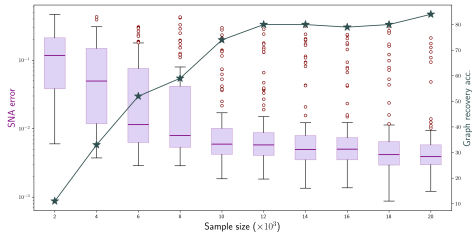
$$\min \|\boldsymbol{\Delta}\|_\infty \quad s.t. \Delta_i = \frac{\left\|\operatorname{proj}_{\operatorname{span}\langle \boldsymbol{h}_j : j \in \overline{\operatorname{sur}}_{\mathcal{G}}(i)\rangle}(\hat{\hat{\boldsymbol{h}}}_i)\right\|_2}{\left\|\hat{\hat{\boldsymbol{h}}}_i\right\|_2}, \tag{5}$$

$$\hat{\hat{\boldsymbol{H}}} = \boldsymbol{P}\hat{\boldsymbol{H}} \text{ for some signed permutation matrix } \boldsymbol{P}.$$

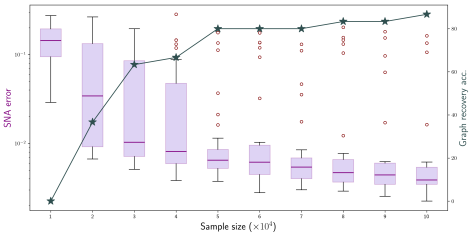
where signed permutation is allowed here since the noise distribution in our experiments is symmetric and the order of latent variables $\boldsymbol{z}_i, i = 1, 2, \cdots, d$ does not matter. We refer to the errors Δ_i defined in (5) as the *SNA error*. The SNA error measures how much of the row $\hat{\hat{\boldsymbol{h}}}_i$ that we learn is contained in the span of the ground-truth rows $\boldsymbol{h}_j, j \in \overline{\operatorname{sur}}_{\mathcal{G}}(i)$. Indeed, recall that given any observation $\boldsymbol{x}$, the ground-truth latent variable is $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$ while our algorithm outputs $\hat{\boldsymbol{v}}_i = \hat{\hat{\boldsymbol{h}}}_i^\top \boldsymbol{x}$, so the SNA error essentially captures whether the recovered latent variable is close to some linear mixture of latent variables in the effect-dominating set of i. When the SNA error is zero for some node i, we know that the recovered latent variable at node i is exactly a linear mixture of the ground-truth latent variables in $\overline{\operatorname{sur}}_{\mathcal{G}}(i)$, according to Lemma 1.



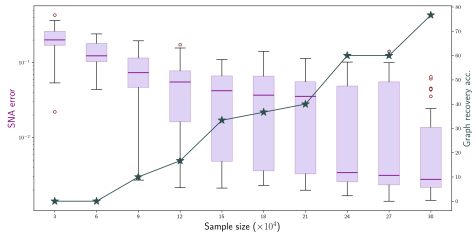
(a) $d = K = 5$



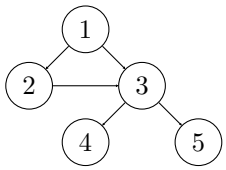
(b) $d = 5, K = 20$



(c) $d = K = 8$



(d) $d = K = 10$



(e) An example causal graph in our experiment

i	$\mathrm{sur}_{\mathcal{G}}(i)$	SNA error	True error
1	$\emptyset$	2.8e-3	2.8e-3
2	{1}	4.0e-3	9.6e-2
3	$\emptyset$	3.4e-3	3.4e-3
4	{3}	1.1e-2	0.99
5	{3}	1.4e-2	0.41

(f) Result for identifying Figure 2e by running `LiNGCReL`

Figure 2: *First two rows*: plots of SNA Error and graph recovery accuracy achieved by `LiNGCReL` as functions of sample size (per environment) for different choices of graph size d and number of environments K. *Third row*: an example of causal graph generated in our experiments, and the estimation error of `LiNGCReL` for each node.

We also define the *true error* for estimating each latent variable. Formally, let $\hat{\boldsymbol{H}}$ be the unmixing matrix that corresponds to the solution of (5), then we define the true estimation error $\tilde{\Delta}_i$ of $\boldsymbol{z}_i$ as

$$\tilde{\Delta}_i = \left\| \left(\boldsymbol{I} - \boldsymbol{h}_i \boldsymbol{h}_i^\top \right) \hat{\boldsymbol{h}}_i \right\|_2 . \tag{6}$$

Results. We randomly sample 100 causal models with size $d = 5$, 30 causal models with size $d = 8$ ad 30 causal models of size $d = 10$. In light of Theorem 1, for each $d \in \{5, 8, 10\}$, we sample data from $K = d$ randomly chosen environments; for $d = 5$ we also consider $K = 20$ to study how different choices of K can affect the result. We run `LiNGCReL` for each model with different sample sizes, compute the SNA error and true error of the obtained solution from (5) and (6) respectively for each latent variable, and check whether the ground-truth causal graph is exactly recovered.

Figure 2 shows how the average SNA error (over all latent variables) and the accuracy of graph recovery changes when sample size grows. We can see `LiNGCReL` successfully recovers about 80% of all models within each category, and the median of the average SNA error is smaller than 1%. Moreover, by comparing Figure 2a with Figure 2b, one can observe that if we fix the total number of samples but choose a larger K (*i.e.*, fewer samples per environment), `LiNGCReL` can still achieve the same level of performance compared with the choice $K = d$. Intuitively, this is because $K \gg d$ vectors sampled from an $r (r \leqslant d)$ dimensional subspace are unlikely to approximately lie in an $(r-1)$-dimensional subspace, so that the calculation of line 6 of Algorithm 2 and line 8 of Algorithm 3 can be more accurate. We leave a better and quantitative understanding of the trade-off between d and K to future work.

SNA error v.s. true error. To understand the implication of our theory, we dive deeper by looking into the learning outcome of `LiNGCReL` on a specific model, of which the causal graph is shown in

Figure 2e. In Figure 2f, we list the surrounding set of each node and the corresponding SNA error and true error. We can see that if $\mathrm{sur}_{\mathcal{G}}(i) = \emptyset$, the two errors equal and both are small, but if $\mathrm{sur}_{\mathcal{G}}(i) \neq \emptyset$, the true error is much larger than the SNA error. This indicates that `LiNGCReL` indeed learns the ground-truth model up to $\sim_{\mathrm{sur}}$, as Theorem 1 predicts.

7 Conclusions

This paper studies the limit of learning identifiable causal representations using data from multiple environments. When hard interventions are not available, we provide theory and algorithm for identification up to SNA, and also show that SNA is an intrinsic ambiguity in our setting.

It is interesting to further investigate the setting where we do not assume that the causal model is linear. Moreover, it is important to understand the concrete form of available interventions in real-world applications. For instance, it is suggested that for single-cell genomics, the intervention is sometimes a "mixture" of hard and soft interventions, and sometimes can even reverse the direction of an edge [43]. Modelling such more complicated interventions appears to be crucial to reveal the underlying causal mechanisms in real-world problems.

Acknowledgments and Disclosure of Funding

VS is supported by NSF Award IIS-2337916 and a 2023 Google Research Scholar Award.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023.

[2] Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. *arXiv preprint arXiv:2310.02854*, 2023.

[3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[5] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[7] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

[8] John Campbell. An interventionist approach to causation in psychology. *Causal learning: Psychology, philosophy, and computation*, pages 58–66, 2007.

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

[10] Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125, 1999.

[11] Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *The Twelfth International Conference on Learning Representations*, 2023.

[12] Kevin N Dunbar and Jonathan A Fugelsang. Causal thinking in science: How scientists and students interpret the unexpected. In *Scientific and technological thinking*, pages 57–79. Psychology Press, 2004.

[13] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 161–168, 2008.

[14] Frederick Eberhardt. Direct causes and the trouble with soft interventions. *Erkenntnis*, 79: 755–777, 2014.

[15] Markus I Eronen. Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59:100785, 2020.

[16] Ronald Aylmer Fisher et al. The design of experiments. *The design of experiments.*, (7th Ed), 1960.

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[18] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[19] Keith J Holyoak and Patricia W Cheng. Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, 62:135–163, 2011.

[20] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

[21] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[23] Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *arXiv preprint arXiv:2305.17225*, 2023.

[24] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. *arXiv preprint arXiv:2306.09643*, 2023.

[25] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[26] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *The Journal of Machine Learning Research*, 21(1):8629–8690, 2020.

[27] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[28] Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023.

[29] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[30] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[31] Judea Pearl. *Causality*. Cambridge university press, 2009.

[32] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

[33] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

[34] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[35] J Schwartz. The formula for change in variables in a multiple integral. *The American Mathematical Monthly*, 61(2):81–85, 1954.

[36] Anna Seigal, Chandler Squires, and Caroline Uhler. Linear causal disentanglement via interventions. *arXiv preprint arXiv:2211.16467*, 2022.

[37] David R Shanks and Anthony Dickinson. Associative accounts of causality judgment. In *Psychology of learning and motivation*, volume 21, pages 229–261. Elsevier, 1988.

[38] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[39] Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

[40] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[41] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. 2000.

[42] Michael Strevens. Review of woodward," making things happen", 2007.

[43] Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

[44] Robert E Tillman and Frederick Eberhardt. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1):41–64, 2014.

[45] Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pages 863–869. Citeseer, 2001.

[46] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

[47] Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. *arXiv preprint arXiv:2310.15450*, 2023.

[48] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021.

[49] Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *arXiv preprint arXiv:2306.00542*, 2023.

[50] Tony Tong Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, et al. Adversarial policies beat superhuman go ais. 2023.

[51] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[52] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[53] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022.

[54] Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.

A Related works

The interventionist approach to causation For the problem of causal graph discovery, it is well-known that the underlying causal structure is non-identifiable given only "passively observed" (equivalently, *i.i.d.*) data alone. As a result, randomized controlled experiments [16] is often used to infer causality. These experiments typically take the form of interventions [41, 31], *i.e.*, manipulations on the "natural state" of the system of interest. Early works [51, 42] define the "hard" (also called "surgical" or "arrow-breaking") interventions in which the value of the intervened variable is entirely determined by the experimenter, thereby removing the dependence of this variable on its direct causes. This type of intervention is arguably the most natural one to consider, and following this definition, a line of works explore sufficient conditions for designing experiments that guarantee identifiability of the causal model in various settings [10, 45, 13, 20, 18].

Intervention *v.s.* passive observation While extensive works demonstrate the success of the interventionist approach, it faces several key challenges that significantly limit its applicability. First, Eberhardt [14] finds that in the presence of unobserved variables, certain causal structures are indistinguishable if we only perform hard interventions. This issue can be resolved by performing soft interventions *i.e.*, interventions that do not remove the dependency on direct causes but only changes the conditional distribution. Second, as pointed out in [44], interventions — whether hard or soft — are often expensive or even infeasible to perform in practice. For example, a psychological intervention is likely to affect multiple psychological variables simultaneously Eronen [15]. As a result, [44] returns to the "passive observation" setting but with multiple datasets with overlapping latent variables.

Interventional causal representation learning Motivated by the interventionist literature in causal graph discovery, a recent line of works [1, 36, 46, 49, 7, 54, 47] consider performing interventions to resolve the non-identifiability issue in causal representation learning [25]. Roughly speaking, these result indicate that identification (possibly with some ambiguities) is possible if one can perform intervention on every latent variable. However, it is unclear how to perform such interventions in practice, given that the underlying latent variables are unknown. Khemakhem et al. [21], Lu et al. [29], Roeder et al. [32] do not require single-node interventions to achieve identifiability, but assumes that the joint distribution of latent variables in each environment lie in a certain exponential family. This assumption can be understood as a prior on the latent variables, but it is unclear when or why it is reasonable to make in reality. Recently, Ahuja et al. [2] considers learning causal representations from multiple domains that relate to each other via an invariance constraint on the subset $\mathcal{S}$ of *stable* latent variables, and they prove identification up to affine mixtures within $\mathcal{S}$.

B Experiment details for Section 6

B.1 Details for step 1 in Section 5

Since $\epsilon = \boldsymbol{B}_k \boldsymbol{z} = \boldsymbol{B}_k \boldsymbol{H} \boldsymbol{x}$ in the k-th environment, so we can use any identification algorithm for linear ICA to recover the matrix $\boldsymbol{M}_k$. Note that while standard linear ICA algorithms only apply to the case where $n = d$, for $n > d$ we can arbitrarily choose d principal components of $\boldsymbol{x}$ to reduce it to the $n = d$ case. This is without loss of generality, since when $n > d$ there is redundant information in $\boldsymbol{x}$.

After recovering $\boldsymbol{M}_k$ for each k by running linear ICA, we still do not know whether each $\boldsymbol{M}_k \boldsymbol{x}$ corresponds to the same permutation of the ground-truth noise variables ϵ. To resolve this issue, we choose test function Ψ mapping any distribution on $\mathbb{R}$ to a deterministic real value, which we expect to take different values for different ϵ_i's. We choose $\Psi(\mathbb{P}) = \mathbb{P}\left[|X| \leqslant 1\right]$ in our experiments. For all $k \geqslant 2$, we calculate the Ψ value of each component of the d-dimensional empirical distribution $\hat{\mathbb{P}}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\boldsymbol{M}_k \boldsymbol{x}_i^{(k)}}$, and choose a permutation π_k to rearrange them in increasing order. Then, we rearrange the columns of $\boldsymbol{M}_k$ using the same permutation π_k. This procedure would asymptotically produce correct alignments as long as $\Psi(\epsilon_i), i \in [d]$ are different, and we find that it empirically works well.

Alternatively, this alignment step can be done as follows: for each pair of environments (E_1, E_t), and for each pair of nodes (i, j), we calculate the distribution distance between ϵ_i in environment E_1 and ϵ_j in environment E_t, based on some notion of distribution distance (*e.g.* kernel maximum

mean discrepancy). Then we find the min-cost perfect matching, where the cost of an edge is the distribution distance.

B.2 Details for the implementation of `LiNGCReL` in the finite-sample regime

Although `LiNGCReL` provably works in the population regime, it faces several challenges when there is only a finite number of samples:

- First, since rank is not a continuous function, it is sensitive to finite-sample estimation errors. In our implementation of Algorithm 3, in each iteration we instead choose $i \notin S$ that has the largest ratio between the first and second singular values of $[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_K]$. And in line 6 of Algorithm 2, we introduce a hyper-parameter `tl` such that the matrix $[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_K]$ is considered to have rank $r_{m'-1}$ if its $r_{m'}$-th singular value is smaller than `tl`. Since the smallest singular value of a random matrix $\boldsymbol{A} \in \mathbb{R}^{K\times m}(K \geqslant m)$ is at the order of $\sqrt{K} - \sqrt{m-1}$ with high probability [33], when $K = d$ one shall choose $\mathrm{tl} \sim \sqrt{d} - \sqrt{d-1} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$. On the other hand, for larger K we can correspondingly choose a larger `tl`. Note that a small `tl` potentially has the risk of being dominated the noise in the estimation, which means that we need more samples per environment to reduce the noise. In contrast, for larger `tl` the estimation is more robust to noise and we can use fewer samples.
- Second, finite-sample estimation errors of $\boldsymbol{M}_k$ make it harder to obtain $\boldsymbol{h}_i$ in Algorithm 3 of Algorithm 3. We implement this step in the following way: first let $\boldsymbol{Q}_j$ be the orthogonal projection matrix onto $E_j^{\perp}$ *i.e.*, $\boldsymbol{Q}_j^{\top}\boldsymbol{x} = \mathrm{proj}_{E_j^{\perp}}(\boldsymbol{x})$, then choose $\boldsymbol{h}_i$ to be the singular vector of $\sum_{j:(j,i)\in\mathcal{E} \text{ or } j=i} \boldsymbol{Q}_j^{\top}\boldsymbol{Q}_j$ that corresponds to the smallest singular value (including zero). Indeed, in the noiseless case we would have $\left(\sum_{j:(j,i)\in\mathcal{E} \text{ or } j=i} \boldsymbol{Q}_j^{\top}\boldsymbol{Q}_j\right)\boldsymbol{h}_i = 0$ if and only if $\boldsymbol{h}_i \in \left(\cap_{j:(i,j)\in\mathcal{E}} E_j\right) \cap E_i$.

C Further experiment results

SNA error v.s. true error We plot the SNA error v.s. true error achieved by `LiNGCReL` in Figure 3. We observe that

- For most nodes, SNA error is exactly equal to the true error and both errors are small, indicating that the corresponding latent variables have been successfully learned by `LiNGCReL`.
- The remaining nodes typically have true error much larger than SNA error. This indicates that there exists some ambiguities at these nodes in the sense that $\mathrm{sur}_{\mathcal{G}}(i) \neq \emptyset$. Note that the true error for many nodes are close to 1; one possible reason is that one selects the wrong singular vector in the second part of Appendix B.2, so that it is orthogonal to the ground-truth vector.

Sensitivity of `LiNGCReL` to the hyperparameter `tl` We examine how different choices of `tl` would affect the performance of `LiNGCReL`. Specifically, we run `LiNGCReL` on the 100 models with size $d = 5$ and number of environments $K = 5$ sampled in Section 6 with $\texttt{tl} \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$ and the results are reported in Figure 4. We can see that the permance is actually quite sensitive to `tl`.

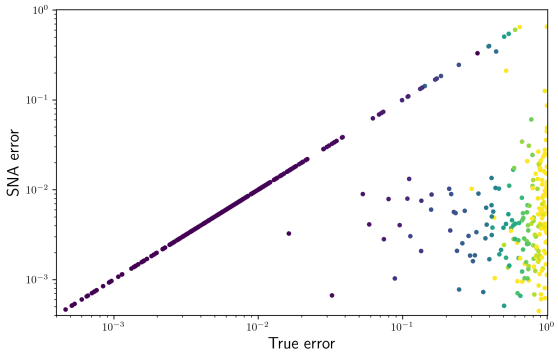


Figure 3: Comparing SNA error with true error for the 500 latent variables in the 100 graphs of size $d = 5$ that we sample in Section 6.

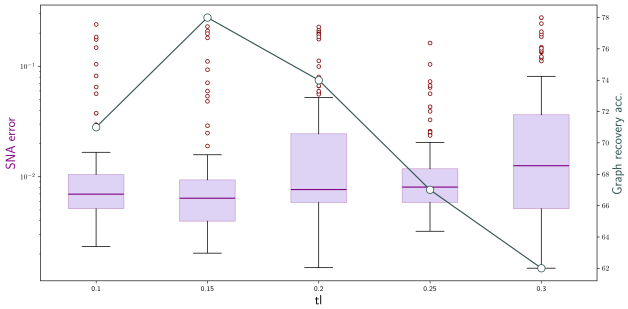


Figure 4: Performance of `LiNGCReL` as a function of `tl`. `tl`$= 0.15$ achieves the best performance in terms of both SNA error and graph recovery accuracy.

D Background on causal representation learning

It is common to assume some axioms on what kind of (conditional) dependency information is encoded in a causal graph (see 41, Section 3.4 for a detailed discussion). The most natural one is the Causal Markov Condition introduced in Definition 1 that gives sufficient conditions for conditional independence via d-separation. We introduce the formal definition of d-separation below:

Definition 5 (paths and colliders)**.** *Let i, j be two nodes of a DAG $\mathcal{G}$, a path is a sequence of nodes $i_0 = i, i_1, \cdots, i_k = j$ such that there is an edge (in either direction) between i_j and $i_{j+1}, j = 0, 1, \cdots, k-1$. A node i_j is called a collider on this path if $i_j \in \mathrm{ch}_{\mathcal{G}}(i_{j-1}) \cap \mathrm{ch}_{\mathcal{G}}(i_{j+1})$.*

Definition 6 (blocked path)**.** *A path in a DAG $\mathcal{G}$ between node i and node j is said to be blocked by a node set S if either of the following holds:*

- *there exists a node v on the path that is in S but not a collider, or*
- *there exists a node v on the path that is a collider, but none of its descendants (including itself) are in S.*

Definition 7 (d-separation)**.** *For a DAG $\mathcal{G}$ with node set $[d]$, any two nodes $i \neq j$ are said to be d-separated by a set $S \subset [d] \setminus \{i, j\}$ if all paths from i to j are blocked by S.*

The *minimality* condition states that there is no redundant edges in the causal graph, and is a natural consequence of the Occam's Razor Principle.

Assumption 6 (Causal minimality, 41, Section 3.4.2)**.** *For latent variables $\boldsymbol{z}$, removing any edge from $\mathcal{G}$ would render violation of the causal Markov condition Definition 1. In other words, let $\mathcal{G}_1$ be the graph obtained by removing any single edge from $\mathcal{G}$, then there must exist $i \in [d]$ such that $\boldsymbol{z}_i \not\perp\!\!\!\perp \boldsymbol{z}_{\mathrm{nd}_{\mathcal{G}_1}(i)} \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}_1}(i)}$.*

The *faithfulness* condition states that the Causal Markov Condition actually entails all (conditional) independence in the latent variables.

Assumption 7 (Faithfulness, 41, Section 3.4.3)**.** *Every (conditional) independence in the latent variables $\boldsymbol{z}$ is entailed by the Causal Markov Condition applied to $\mathcal{G}$. In other words, $\boldsymbol{z}_i \perp\!\!\!\perp \boldsymbol{z}_j \mid \boldsymbol{z}_S \Leftrightarrow i, j$ are d-separated by S.*

Existing works have explored different notions of identifiability. For observational data, it is well known that Markov equivalence of graphs is an intrinsic ambiguity that one cannot resolve:

Definition 8 (Markov equivalence/Faithful Indistinguishability, 41, Section 4.2)**.** *If two DAGs encodes the same set of dependency relations, we say that they are Markov equivalent.*

Any DAG $\mathcal{G}$ induces a partial order on its nodes which we denote by $\prec_{\mathcal{G}}$. In the special case when for all $i \neq j$, either $i \prec_{\mathcal{G}} j$ or $j \prec_{\mathcal{G}} i$ holds, we say that $\prec_{\mathcal{G}}$ is a total order. This partial order is equivalent to the transitional closure of the graph, as defined below:

Definition 9 (Transitional closure)**.** *Given any DAG $\mathcal{G}$, its transitional closure $\bar{\mathcal{G}}$ is defined to be the graph obtained by connecting all edges $i \to j$ where i is an ancestor of j in $\mathcal{G}$.*

When $\prec_{\mathcal{G}}$ is a total order, each pair of nodes are connected by a directed edge in its transitive closure $\bar{\mathcal{G}}$. Such $\bar{\mathcal{G}}$ is often called a *tournament* in graph theory.

In the following, we list different forms of identifiability that appear in the literature:

Definition 10 (different notions of identifiability)**.** *Let $\mathcal{H} : \mathbb{R}^n \supseteq \mathcal{X} \mapsto \mathbb{R}^d$ be the space of diffeomorphic mappings from observation to latent, and $\mathfrak{G}$ be the space of all DAGs with d nodes, then for $h, \hat{h} \in \mathcal{H}$ and $\mathcal{G}, \hat{\mathcal{G}} \in \mathfrak{G}$, we write*

(i) [36, 23] $(h, \mathcal{G}) \overset{T}{\sim}_G (\hat{h}, \hat{\mathcal{G}})$ if there exists a permutation π on $[d]$ such that $\pi(\mathcal{G})$ and $\hat{\mathcal{G}}$ have the same transitional closure;

(ii) [49, 47] $(h, \mathcal{G}) \sim_{\mathrm{CRL}} (\hat{h}, \hat{\mathcal{G}})$ if we actually have $\mathcal{G} = \hat{\mathcal{G}}$ for the ϕ defined above.

Given an equivalence relation $\sim$ on $\mathcal{H} \times \mathfrak{G}$, we say that a causal model $(\boldsymbol{h}, \mathcal{G})$ is identifiable under $\sim$ if any candidate solution $(\hat{\boldsymbol{h}}, \hat{\mathcal{G}})$ satisfies $(\hat{\boldsymbol{h}}, \hat{\mathcal{G}}) \sim (\boldsymbol{h}, \mathcal{G})$. The notion of identification up to $\overset{T}{\sim}_G$, as shown in Seigal et al. [36] with single-node soft interventions on linear causal models, is highly related to this paper. Compared with their result, our $\sim_{\mathrm{sur}}$ guarantee is must stronger, since not only the causal graph can be fully recovered, but the latent variables can be identified up to mixtures of the effect-dominating sets as well.

E Illustrating examples for our theory and algorithm

E.1 An example for understanding the SNA ambiguity

We provide a simple example below to illustrate the SNA ambiguity discussed in Section 3.

Example 1. *Let G be a causal graph with $d = 3$ nodes and edges $1 \to 2$ and $2 \to 3$. We have access to observations from a set of environments $\mathfrak{E}$. It turns out that there is no way to distinguish between the following two structural equation models:*

$$
\begin{aligned}
\boldsymbol{z}_1 &= \epsilon_1^E \\
\boldsymbol{z}_2 &= f_2^E(\boldsymbol{z}_1, \epsilon_2^E) \\
\boldsymbol{z}_3 &= f_3^E(\boldsymbol{z}_2, \epsilon_3^E) \\
\boldsymbol{x} &= \boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)^\top
\end{aligned}
\qquad\qquad
\begin{aligned}
\boldsymbol{v}_1 &= \epsilon_1^E \\
\boldsymbol{v}_2 &= f_2^E(\boldsymbol{v}_1, \epsilon_2^E) \\
\boldsymbol{v}_3 &= \boldsymbol{v}_2 + f_3^E(\boldsymbol{v}_2, \epsilon_3^E) \\
\boldsymbol{x} &= (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 - \boldsymbol{v}_2)^\top
\end{aligned}
$$

where $\epsilon_i^E, i = 1, 2, 3$ are independent noise variables, if we do not change the causal graph $\mathcal{G}$, no matter what environment E that we have.

This issue does not exist when we assume access to hard interventions on node 3, which effectively removes the edge $2 \to 3$. Specifically, with hard intervention on $\boldsymbol{z}_3$, the variables $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$ become independent. But by definition, $\boldsymbol{v}_2 = \boldsymbol{z}_2$ and $\boldsymbol{v}_3 = \boldsymbol{z}_2 + \boldsymbol{z}_3$ must be dependent, so this intervention

cannot be realized by any hard intervention on $\boldsymbol{v}_3$, thereby providing a way to distinguish between the above models.

Without node 3, the same ambiguity would arise on node 2. However, node 3 can help us to overcome this ambiguity, thanks to the fact that node 2 is the only causal parent of node 3. Suppose for example that $\boldsymbol{v}_2 = m(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is some mixture of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, then $\boldsymbol{v}_3 = \hat{f}_3^E\left(\boldsymbol{v}_2, \epsilon_3^E\right) = \hat{f}_3^E\left(m(\boldsymbol{z}_1, \boldsymbol{z}_2), \epsilon_3^E\right)$. Since all environments share the same mixing function, $\boldsymbol{v}_3$ must be some deterministic function $\psi_3(\boldsymbol{z})$ of $\boldsymbol{z}$, where ψ_3 is the same across all environment E. Hence, we have

$$\hat{f}_3^E\left(m(\boldsymbol{z}_1, \boldsymbol{z}_2), \epsilon_3^E\right) = \psi_3\left(\boldsymbol{z}_1, \boldsymbol{z}_2, f_3^E(\boldsymbol{z}_2, \epsilon_3^E)\right) \tag{7}$$

Now we note that the dependencies of LHS on $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are through a single scalar-valued function m, but since we would have different f_3^E's in different environments, this in general does not hold for the RHS. Therefore, any causal model with latent variable $\boldsymbol{v}_2$ as a mixture of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ cannot be equivalent to the ground-truth model.

According to Definition 3, in Example 1 we have $\mathrm{sur}_{\mathcal{G}}(1) = \mathrm{sur}_{\mathcal{G}}(2) = \emptyset$ but $\mathrm{sur}_{\mathcal{G}}(3) = \{2\}$.

E.2 An example for the main idea behind LiNGCReL

To illustrate our main algorithm on how we can recover the graph $\mathcal{G}$ and the matrix $\boldsymbol{H}$, we first provide some intuition using a simple three-node example:

Example 2. *Let $\mathcal{G}$ be the graph with $d = 3$ nodes and edges $1 \to 2, 1 \to 3$ and $2 \to 3$, so that each $\boldsymbol{B}_k$ is of form*

$$\boldsymbol{B}_k = \begin{pmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ \times & \times & \times \end{pmatrix} \begin{matrix} \rightsquigarrow \boldsymbol{b}_{k1} \\ \rightsquigarrow \boldsymbol{b}_{k2} \\ \rightsquigarrow \boldsymbol{b}_{k3} \end{matrix} \tag{8}$$

*We can identify the graph as follows: first, for $i \in \{1, 2, 3\}$, look at the space $\boldsymbol{W}_i$ spanned by the rows $(\boldsymbol{M}_k)_i, k \in [K]$. If $\dim \boldsymbol{W}_i = 1$, we know that i is a source node (*i.e.*, $\mathrm{pa}_{\mathcal{G}}(i) = \emptyset$) in $\mathcal{G}$. Otherwise it is not, due to Assumption 5. Hence we can know that node 1 is a source node.*

In our example, there is no other node that satisfies this requirement. We then proceed to search for some $i \neq 1$ such that the projection of $\boldsymbol{W}_i$ onto $\boldsymbol{W}_1^{\perp}$ has dimension 1. If this holds, then one can show that $\mathrm{pa}_{\mathcal{G}}(i) = \{1\}$. Otherwise, i must have parents other than 1.

It turns this requirement is satisfied for node 2 since $\dim\left(\mathrm{proj}_{\boldsymbol{h}_1}\mathrm{span}\left\langle \boldsymbol{h}_1, \boldsymbol{h}_2\right\rangle\right) = 1$, but is not satisfied for node 3 since $\dim\left(\mathrm{proj}_{\boldsymbol{h}_1}\mathrm{span}\left\langle \boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3\right\rangle\right) \geqslant 2$ (by Lemma 4). Hence we know that $\mathrm{pa}_{\mathcal{G}}(2) = \{1\}$.

Finally, it remains to determine $\mathrm{pa}_{\mathcal{G}}(3)$. To do this, we first note that $\dim \boldsymbol{W}_3 = 3$. Then we project $\boldsymbol{W}_3$ onto $\boldsymbol{W}_1^{\perp}$ and $\boldsymbol{W}_2^{\perp}$ respectively, and the resulting dimensions are 2 and 1. As we rigorously show in Proposition 2, a decrease of the dimension exactly indicates finding a new parent. Thus we have $\mathrm{pa}_{\mathcal{G}}(3) = \{1, 2\}$, completing the recovery of the graph.

Finally, we recover the unmixing matrix $\boldsymbol{H}$ (and thus the latent variables) by noticing that $\boldsymbol{h}_1 \in \boldsymbol{W}_1$, $\boldsymbol{h}_2 \in \boldsymbol{W}_2 \cap \boldsymbol{W}_3$ and $\boldsymbol{h}_3 \in \boldsymbol{W}_3$. Ambiguities would arise at nodes 2 and 3, which are exactly the nodes that have non-empty effect-dominating sets.

F Auxiliary lemmas

Lemma 2. *For any family of m-dimensional vectors $\{\boldsymbol{v}_k\}_{k=1}^K$ and $\{\boldsymbol{z}_k\}_{k=1}^K$ if $\boldsymbol{v}_k = \boldsymbol{z}_k\boldsymbol{T}$ and $\boldsymbol{T} \in \mathbb{R}^{m\times m}$ is invertible, then*

$$\dim \mathrm{span}\left\langle \boldsymbol{v}_k : k \in [K]\right\rangle = \dim \mathrm{span}\left\langle \boldsymbol{z}_k : k \in [K]\right\rangle$$

Theorem 5 (Darmois-Skitovic Theorem)**.** *Let $\epsilon_i, i \in [d]$ be independent random variables and $X = \sum_{i=1}^d \alpha_i\epsilon_i, Y = \sum_{i=1}^d \beta_i\epsilon_i$. If $X \perp\!\!\!\perp Y$, then for $\forall i \in [d]$, $\alpha_i\beta_i \neq 0 \Rightarrow \epsilon_i$ is Gaussian distributed.*

Lemma 3. *Suppose that $\epsilon = (\epsilon_1, \cdots, \epsilon_d)$ is a d-dimensional random vector with independent components such that $\mathrm{Var}(\epsilon_i) = 1, \forall i \in [d]$, and there exists an invertible and non-diagonal matrix $\boldsymbol{M}$ such that $\boldsymbol{M}\epsilon \overset{d}{=} \epsilon$, then at least one of the following statements must hold:*

(1) there exists at least two Gaussian variables in $\epsilon_1, \cdots, \epsilon_d$;

(2) $\boldsymbol{M}$ is a permutation matrix and there exists $1 \leqslant i < j \leqslant d$ such that $\epsilon_i \overset{d}{=} \epsilon_j$.

Proof. Suppose that (1) does not hold, then there is at most one Gaussian variable in $\epsilon_1, \cdots, \epsilon_d$. We assume WLOG that $\epsilon_1, \cdots, \epsilon_{d-1}$ are all non-Gaussian. Then by the Darmois-Skitovic Theorem, we know that for $\forall 1 \leqslant j < k \leqslant [d]$ and $i \in [d-1]$, $\boldsymbol{M}_{ji} \cdot \boldsymbol{M}_{ki} = 0 \Rightarrow$ there is at most one non-zero entry in each of the first $d-1$ columns of $\boldsymbol{M}$.

Assume that $\boldsymbol{M}_{k_i,i} \neq 0$, $i \in [d-1]$. Since $\boldsymbol{M}$ is invertible, we know that $k_i, i \in [d-1]$ must be different. Let k_d be the remaining element in $[d]$ that does not appear in $k_i, i < d$, then $(\boldsymbol{M}\epsilon)_{k_d} = \boldsymbol{M}_{k_d,d}\epsilon_d$, while $(\boldsymbol{M}\epsilon)_{k_i} = \boldsymbol{M}_{k_i,i}\epsilon_i + \boldsymbol{M}_{k_i,d}\epsilon_d$. Since the components of $\boldsymbol{M}\epsilon$ are independent, it is easy to see that $\boldsymbol{M}_{id} \neq 0, \forall i \neq k_d$. In other words, $\boldsymbol{M}$ only has non-zero entries at $(k_i, i), i \in [d]$.

Since $\mathrm{Var}(\epsilon_i) = 1$, we know that $\boldsymbol{M}$ must be a signed permutation matrix. Finally, let π be the permutation on $[d]$ such that $\boldsymbol{M}_{i,\pi(i)} \neq 0$. Since $\boldsymbol{M}$ is not diagonal, π must have a cycle $(i_1, i_2, \cdots, i_k)$ with length $k \geqslant 2$, so that $\epsilon_{i_1}, \cdots, \epsilon_{i_k}$ all have the same distribution, which implies that (2) holds, as desired. □

Lemma 4. *Let $\boldsymbol{V}_1, \boldsymbol{V}_2$ be two subspaces of $\mathbb{R}^d$ such that $\boldsymbol{V}_1 \cap \boldsymbol{V}_2 = \{\boldsymbol{0}\}$, and $\boldsymbol{P}_{\boldsymbol{V}_1^\perp}$ be the orthogonal projection onto $\boldsymbol{V}_1^\perp$, then we have that $\dim(\boldsymbol{V}_2) = \dim\left(\boldsymbol{P}_{\boldsymbol{V}_1^\perp}\boldsymbol{V}_2\right)$.*

Proof. Obviously we have $\dim(\boldsymbol{V}_2) \geqslant \dim\left(\boldsymbol{P}_{\boldsymbol{V}_1^\perp}\boldsymbol{V}_2\right)$. On the other hand, let $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m$ be a basis of $\boldsymbol{V}_2$, then $\boldsymbol{w}_i = \boldsymbol{P}_{\boldsymbol{V}_1^\perp}\boldsymbol{u}_i, i = 1, 2, \cdots, m$ are also independent. Indeed, suppose that $\lambda_i, i = 1, 2, \cdots, m$ satisfy $\sum_{i=1}^m \lambda_i \boldsymbol{w}_i = 0$, then $\boldsymbol{P}_{\boldsymbol{V}_1^\perp}\left(\sum_{i=1}^m \lambda_i \boldsymbol{u}_i\right) = 0$, implying that $\sum_{i=1}^m \lambda_i \boldsymbol{u}_i \in \boldsymbol{V}_1$. However, we know that $\boldsymbol{V}_1 \cap \boldsymbol{V}_2 = \{\boldsymbol{0}\}$, so $\lambda_1 = \cdots = \lambda_m = 0$. This concludes the proof. □

Lemma 5. *Assumption 4 is equivalent to Assumption 5.*

Proof. The main observation is that for each $k \in [K]$, $(\boldsymbol{B}_k)_i$ only has non-zero entries at the j-th coordinate where $j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$. Moreover, let $\tilde{\boldsymbol{w}}_k(i)$ be the vector consisting of these entries, then $\tilde{\boldsymbol{w}}_k(i) = (\boldsymbol{\Omega}_k)_{ii}^{-\frac{1}{2}}(-\boldsymbol{w}_k(i), 1)$. Hence,

$$\dim \operatorname{span}\langle (\boldsymbol{B}_k)_i : k \in [K] \rangle = \dim \operatorname{span}\langle (-\boldsymbol{w}_k(i), 1) : k \in [K] \rangle .$$

Suppose that Assumption 4 holds, then for $\forall \boldsymbol{x} \in \mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|}$, there exists $\lambda_k \in \mathbb{R}, 1 \leqslant k \leqslant \left|\mathrm{pa}_{\mathcal{G}}(i)\right|$ such that $\sum_k \lambda_k = 1$ and $\sum_k \lambda_k \boldsymbol{w}_k(i) = \boldsymbol{x}$. Hence,

$$(\boldsymbol{x}, 1) = \sum_k \lambda_k \tilde{\boldsymbol{w}}_k(i) \in \operatorname{span}\langle (\boldsymbol{B}_k)_i : k \in [K] \rangle .$$

This immediately implies that $\operatorname{span}\langle (\boldsymbol{B}_k)_i : k \in [K] \rangle = \mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|+1}$, so that Assumption 5 holds.

Conversely, suppose that Assumption 5 holds, then for $\forall \boldsymbol{x} \in \mathbb{R}^{|\mathrm{pa}_{\mathcal{G}}(i)|}$, there exists $\lambda_k \in \mathbb{R}, 1 \leqslant k \leqslant \left|\mathrm{pa}_{\mathcal{G}}(i)\right|$ such that $\sum_k \lambda_k \tilde{\boldsymbol{w}}_k(i) = (\boldsymbol{x}, 1)$. Hence we have $\sum_k \lambda_k \boldsymbol{w}_k(i) = \boldsymbol{x}$ and $\sum_k \lambda_k = 1$, implying Assumption 4. □

G Properties of effect-domination sets

Lemma 6.

- *$j \in \mathrm{sur}_{\mathcal{G}}(i)$ if and only if $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$;*
- *when $i \neq j$, $j \in \mathrm{sur}_{\mathcal{G}}(i)$ if and only if $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \overline{\mathrm{ch}}_{\mathcal{G}}(j)$.*

Proof. If $j \in \mathrm{sur}_{\mathcal{G}}(i)$, by definition $i \in \mathrm{ch}_{\mathcal{G}}(j)$ and $\mathrm{ch}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$, so that $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$. Conversely, $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$ implies that $i \in \mathrm{ch}_{\mathcal{G}}(j)$ and $\mathrm{ch}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$, so $j \in \mathrm{sur}_{\mathcal{G}}(i)$. This proves the first claim.

To prove the second claim, assume that $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \overline{\mathrm{ch}}_{\mathcal{G}}(j)$ holds but $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$ does not hold, then we must have $j \in \overline{\mathrm{ch}}_{\mathcal{G}}(i)$. since $j \neq i$, we have $j \in \mathrm{ch}_{\mathcal{G}}(i)$, but then $i \notin \overline{\mathrm{ch}}_{\mathcal{G}}(j)$, which is a contradiction. Hence $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subseteq \mathrm{ch}_{\mathcal{G}}(j)$ and the conclusion follows from the first claim. □

Lemma 7. *Let $\mathcal{G}$ be a DAG and i be its node, then for $\forall j \in \mathrm{pa}_{\mathcal{G}}(i)$, we have $\mathrm{sur}_{\mathcal{G}}(j) \subseteq \mathrm{pa}_{\mathcal{G}}(i)$.*

Proof. Let $k \in \mathrm{sur}_{\mathcal{G}}(j)$, then by definition we have $\mathrm{ch}_{\mathcal{G}}(j) \subseteq \mathrm{ch}_{\mathcal{G}}(k)$. In particular, we have $i \in \mathrm{ch}_{\mathcal{G}}(k) \Rightarrow k \in \mathrm{pa}_{\mathcal{G}}(i)$. □

Lemma 8. *Let $\mathcal{G}$ be a DAG and i be its node, then for $\forall j \in \mathrm{sur}_{\mathcal{G}}(i)$, we have $\mathrm{sur}_{\mathcal{G}}(j) \subseteq \mathrm{sur}_{\mathcal{G}}(i)$.*

Proof. Let $k \in \mathrm{sur}_{\mathcal{G}}(j)$, then by definition we have $\overline{\mathrm{ch}}_{\mathcal{G}}(j) \subset \overline{\mathrm{ch}}_{\mathcal{G}}(k)$. We also know that $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subset \overline{\mathrm{ch}}_{\mathcal{G}}(j)$, so $\overline{\mathrm{ch}}_{\mathcal{G}}(i) \subset \overline{\mathrm{ch}}_{\mathcal{G}}(k)$, implying that $k \in \mathrm{sur}_{\mathcal{G}}(i)$. □

Lemma 9. *If $\boldsymbol{M} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$, then $\boldsymbol{M}^{-1} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$.*

Proof. Assume WLOG that the nodes of $\mathcal{G}$ satisfy $i \in \mathrm{pa}_{\mathcal{G}}(j) \Rightarrow i < j$ (otherwise we can choose a different index of the nodes and correspondingly swap some rows and columns of $\boldsymbol{M}$). Since $i \in \mathrm{sur}_{\mathcal{G}}(j) \Rightarrow i \in \mathrm{pa}_{\mathcal{G}}(j)$, it follows that $\boldsymbol{M}$ must be lower triangular and the diagonal entries are nonzero.

Let $\boldsymbol{N} = \boldsymbol{M}^{-1}$, then for $\forall i \in [d]$, we have

$$\sum_{j=1}^{d} \boldsymbol{N}_{ij}\boldsymbol{M}_{j\ell} = 0, \quad \forall \ell \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i). \tag{9}$$

Since $\boldsymbol{M} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$, we have $\boldsymbol{M}_{j\ell} = 0$ for $\forall j$ such that $\ell \notin \overline{\mathrm{sur}}_{\mathcal{G}}(j)$. By Lemma 8, if $j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, then $\ell \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$ necessarily implies that $\ell \notin \overline{\mathrm{sur}}_{\mathcal{G}}(j)$. Hence the left hand side of (9) is essentially a sum over $j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, *i.e.*,

$$\sum_{j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)} \boldsymbol{N}_{ij}\boldsymbol{M}_{j\ell} = 0, \quad \forall \ell \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i).$$

Viewing the above as a system of linear equations in $\boldsymbol{N}_{ij}, j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, the coefficient matrix $(\boldsymbol{M}_{j\ell})_{j,\ell \in \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)}$ must be invertible since it is a sub-matrix of the invertible lower-triangular matrix $\boldsymbol{M}$. As a result, we necessary have $\boldsymbol{N}_{ij} = 0, \forall j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$. Finally, $\boldsymbol{N} = \boldsymbol{M}^{-1}$ must be invertible, so $\boldsymbol{N} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$ as desired. □

Lemma 10. *Suppose that $\psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a diffeomorphism and $\mathcal{G}$ be a DAG, such that for $\forall i \in [d]$, $\psi_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\mathrm{sur}}_{\mathcal{G}}(i)}$. Then for $\forall j \in [d]$, $(\psi^{-1})_j(\boldsymbol{v})$ is a function of $\boldsymbol{v}_{\overline{\mathrm{sur}}_{\mathcal{G}}(j)}$.*

Proof. Let $\boldsymbol{J}_{\boldsymbol{z}} = \boldsymbol{J}_{\psi}(\boldsymbol{z})$ be the Jacobian matrix of ψ. Since ψ is a diffeomorphism, $\boldsymbol{J}_{\boldsymbol{z}}$ is invertible for any $\boldsymbol{z} \in \mathbb{R}^d$. Moreover, our assumption implies that $(\boldsymbol{J}_{\boldsymbol{z}})_{ij} = 0, \forall j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, so $\boldsymbol{J}_{\boldsymbol{z}} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$. By Lemma 9, $\boldsymbol{J}_{\boldsymbol{z}}^{-1} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$. But $\boldsymbol{J}_{\boldsymbol{z}}^{-1}$ is exactly the Jacobian matrix of ψ^{-1} at $\boldsymbol{v} = \psi(\boldsymbol{z})$, hence it follows that $(\psi^{-1})_j(\boldsymbol{v})$ is only a function of $\boldsymbol{v}_{\overline{\mathrm{sur}}_{\mathcal{G}}(j)}$, as desired. □

Lemma 11. *The binary relation $\sim_{\mathrm{sur}}$ defined in Definition 4 is an equivalence relation.*

Proof. It is obvious that $(\boldsymbol{h}, \mathcal{G}) \sim_{\mathrm{sur}} (\boldsymbol{h}, \mathcal{G})$ holds for any model $(\boldsymbol{h}, \mathcal{G})$.

Suppose that $(\boldsymbol{h}_1, \mathcal{G}_1) \sim_{\mathrm{sur}} (\boldsymbol{h}_2, \mathcal{G}_2)$, then there exists a permutation π on $[d]$ and a diffeomorphism $\psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ where $\psi_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\mathrm{sur}}_{\mathcal{G}_1}(i)}$, such that $i \in \mathrm{pa}_{\mathcal{G}_1}(j) \Leftrightarrow \pi(i) \in \mathrm{pa}_{\mathcal{G}_2}(\pi(j))$ and $\boldsymbol{P}_{\pi} \circ \boldsymbol{h}_2 = \psi \circ \boldsymbol{h}_1$. Then we can write $\boldsymbol{P}_{\pi}^{-1} \circ \boldsymbol{h}_1 = \hat{\psi} \circ \boldsymbol{h}_2$ where $\hat{\psi} = \boldsymbol{P}_{\pi}^{-1} \circ \psi^{-1} \circ \boldsymbol{P}_{\pi}$. By Lemma 10, we know that $\left(\psi^{-1}\right)_j(\boldsymbol{v})$ is a function of $\boldsymbol{v}_{\overline{\mathrm{sur}}_{\mathcal{G}_1}(j)}$, so $(\hat{\psi})_j$ is a function of $\boldsymbol{v}_{\pi(\overline{\mathrm{sur}}_{\mathcal{G}_1}(j))} = \boldsymbol{v}_{\overline{\mathrm{sur}}_{\mathcal{G}_2}(j)}$, implying that $(\boldsymbol{h}_2, \mathcal{G}_2) \sim_{\mathrm{sur}} (\boldsymbol{h}_1, \mathcal{G}_1)$.

Finally, let $(\boldsymbol{h}_1, \mathcal{G}_1) \sim_{\text{sur}} (\boldsymbol{h}_2, \mathcal{G}_2)$ and $(\boldsymbol{h}_2, \mathcal{G}_2) \sim_{\text{sur}} (\boldsymbol{h}_3, \mathcal{G}_3)$, then we can write

$$\boldsymbol{P}_\pi \circ \boldsymbol{h}_2 = \psi \circ \boldsymbol{h}_1 \quad \text{and} \quad \boldsymbol{P}_{\hat{\pi}} \circ \boldsymbol{h}_3 = \hat{\psi} \circ \boldsymbol{h}_2$$

where: for $\forall i \in [d]$, $\psi_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_1}(i)}$, $\hat{\psi}_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_2}(i)}$, $i \in \text{pa}_{\mathcal{G}_1}(j) \Leftrightarrow \pi(i) \in \text{pa}_{\mathcal{G}_2}(\pi(j))$ and $i \in \text{pa}_{\mathcal{G}_2}(j) \Leftrightarrow \hat{\pi}(i) \in \text{pa}_{\mathcal{G}_2}(\hat{\pi}(j))$. Then, we can write

$$\boldsymbol{P}_\pi \circ \boldsymbol{P}_{\hat{\pi}} \circ \boldsymbol{h}_3 = \boldsymbol{P}_\pi \circ \hat{\psi} \circ \boldsymbol{P}_\pi^{-1} \circ \psi \circ \boldsymbol{h}_1.$$

Since $\hat{\psi}_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_2}(i)}$, we deduce that $\left(\boldsymbol{P}_\pi \circ \hat{\psi} \circ \boldsymbol{P}_\pi^{-1}\right)_i(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_1}(i)}$. Hence, $\left(\boldsymbol{P}_\pi \circ \hat{\psi} \circ \boldsymbol{P}_\pi^{-1} \circ \psi\right)_i(\boldsymbol{z}) = \left(\boldsymbol{P}_\pi \circ \hat{\psi} \circ \boldsymbol{P}_\pi^{-1}\right)_i(\psi(\boldsymbol{z}))$ is a function of $\psi_{\overline{\text{sur}}_{\mathcal{G}_1}(i)}(\boldsymbol{z})$. The definition of ψ implies that for each $j \in \overline{\text{sur}}_{\mathcal{G}_1}(i)$, $\psi_j(\boldsymbol{z})$ is a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_1}(j)}$. By Lemma 8, we have $\cup_{j \in \overline{\text{sur}}_{\mathcal{G}_1}(i)} \overline{\text{sur}}_{\mathcal{G}_1}(j) \subseteq \overline{\text{sur}}_{\mathcal{G}_1}(i)$. Hence $\left(\boldsymbol{P}_\pi \circ \hat{\psi} \circ \boldsymbol{P}_\pi^{-1} \circ \psi\right)_i(\boldsymbol{z})$ is still a function of $\boldsymbol{z}_{\overline{\text{sur}}_{\mathcal{G}_1}(i)}$. Moreover, we also have $i \in \text{pa}_{\mathcal{G}_1}(j) \Leftrightarrow \pi(i) \in \text{pa}_{\mathcal{G}_2}(\pi(j)) \Leftrightarrow \hat{\pi} \circ \pi(i) \in \text{pa}_{\mathcal{G}_2}(\hat{\pi} \circ \pi(j))$, so by definition, $(\boldsymbol{h}_1, \mathcal{G}_1) \sim_{\text{sur}} (\boldsymbol{h}_3, \mathcal{G}_3)$, as desired. □

H Omitted proofs from Section 4 and Section 5

H.1 Proof of Theorem 1

According to the assumption, we have that $\epsilon = \boldsymbol{B}_k \boldsymbol{H} \boldsymbol{x}$ and $\hat{\epsilon} = \hat{\boldsymbol{B}}_k \hat{\boldsymbol{H}} \boldsymbol{x}$, so that $\epsilon = \boldsymbol{B}_k \boldsymbol{H} (\hat{\boldsymbol{B}}_k \hat{\boldsymbol{H}})^\dagger \hat{\epsilon}, \forall k \in [K]$. By Lemma 3, we know that for each k, $\boldsymbol{P}_k := \boldsymbol{B}_k \boldsymbol{H} (\hat{\boldsymbol{B}}_k \hat{\boldsymbol{H}})^\dagger$ is a signed permutation matrix, so that $\epsilon = \boldsymbol{P}_k \hat{\epsilon}$. Since for any $i \neq j$, $\hat{\epsilon}_i \overset{d}{\neq} \hat{\epsilon}_j$, we must have $|\boldsymbol{P}|_1 = |\boldsymbol{P}|_2 = \cdots = |\boldsymbol{P}|_K =: \boldsymbol{P}$ and $\epsilon = \boldsymbol{P} \hat{\epsilon}$, where $|M|$ denotes the resulting matrix by taking the absolute value of all entries in $\boldsymbol{M}$. Thus, we can WLOG assume that $\epsilon = \hat{\epsilon}$, since otherwise we can permute the noise variables $\hat{\epsilon}$, and also permute the rows of $\boldsymbol{B}_k$ correspondingly. In other words, suppose that the permutation matrix $|\boldsymbol{P}|$ has $|\boldsymbol{P}|_{k_i, i} = 1, i \in [d]$, then we can assign to each node i in $\hat{\mathcal{G}}$ a new index k_i and work with the new indices.

In this case, by Lemma 3 we have $\boldsymbol{B}_k \boldsymbol{H} = \boldsymbol{\Sigma}_k \hat{\boldsymbol{B}}_k \hat{\boldsymbol{H}}, \forall k \in [K]$ or equivalently $\boldsymbol{\Sigma}_k \hat{\boldsymbol{B}}_k = \boldsymbol{B}_k \boldsymbol{T}$, where $\boldsymbol{T} = \boldsymbol{H} \hat{\boldsymbol{H}}^\dagger \in \mathbb{R}^{d \times d}$, and $\boldsymbol{\Sigma}_k$ is a diagonal matrix with diagonal entries in $\{+1, -1\}$. Let $\hat{\hat{\boldsymbol{B}}}_k = \boldsymbol{\Sigma}_k \hat{\boldsymbol{B}}_k$, then the rows of $\hat{\hat{\boldsymbol{B}}}_k$ equals (up to sign) to the rows of $\hat{\boldsymbol{B}}_k$.

To summarize, we now know that i) $\hat{\hat{\boldsymbol{B}}}_k = \boldsymbol{B}_k \boldsymbol{T}, k \in [K]$, ii) $(\boldsymbol{B}_k)_{ij} \neq 0 \Leftrightarrow j \in \overline{\text{pa}}_{\mathcal{G}}(i)$, and similarly, $(\hat{\hat{\boldsymbol{B}}}_k)_{ij} \neq 0 \Leftrightarrow j \in \overline{\text{pa}}_{\hat{\mathcal{G}}}(i)$, and iii) Both $\{\boldsymbol{B}_k\}$ and $\left\{\hat{\hat{\boldsymbol{B}}}_k\right\}$ satisfy the node-level non-degeneracy assumption Assumption 5. For any two such matrices that satisfy such a set of conditions, it must necessarily be true that $\mathcal{G} = \hat{\mathcal{G}}$.

Lemma 12 (Graph Identifiability)**.** *Consider any two sets matrices $\{\hat{\hat{\boldsymbol{B}}}_k\}_{k \in [K]}$ and $\{\boldsymbol{B}_k\}_{k \in [K]}$ and associated graphs $\mathcal{G}, \hat{\mathcal{G}}$. If these sets and graphs satisfy that:*

1. *$\hat{\hat{\boldsymbol{B}}}_k = \boldsymbol{B}_k \boldsymbol{T}, k \in [K]$;*
2. *$(\boldsymbol{B}_k)_{ij} \neq 0 \Leftrightarrow j \in \overline{\text{pa}}_{\mathcal{G}}(i)$, and similarly, $(\hat{\hat{\boldsymbol{B}}}_k)_{ij} \neq 0 \Leftrightarrow j \in \overline{\text{pa}}_{\hat{\mathcal{G}}}(i)$.*
3. *Both $\{\boldsymbol{B}_k\}$ and $\left\{\hat{\hat{\boldsymbol{B}}}_k\right\}$ satisfy the node-level non-degeneracy assumption Assumption 5.*

then it must hold that $\mathcal{G} = \hat{\mathcal{G}}$.

Proof. We prove this via induction on the size of the graph d. Note that here $\mathcal{G} = \hat{\mathcal{G}}$ is not up to permutation and our statement is equivalent to $\text{pa}_{\mathcal{G}}(i) = \text{pa}_{\hat{\mathcal{G}}}(i), \forall i \in [d]$.

If $d = 1$, *i.e.*, $\mathcal{G} = \hat{\mathcal{G}}$ obviously holds since both are graphs with only 1 node.

Suppose that for all graphs $\mathcal{G}$ of size $d-1$, the graph $\hat{\mathcal{G}}$ satisfying all given assumptions must necessarily be equal to $\mathcal{G}$. Now, we consider the case that $\mathcal{G}$ has d nodes. WLOG we can assume that the nodes of $\mathcal{G}$ are properly indexed such that $i \in \mathrm{pa}_{\mathcal{G}}(j) \Rightarrow i < j$, so $\boldsymbol{B}_k, k \in [K]$ are lower-triangular matrices. (However, it is currently unknown whether $\hat{\hat{\boldsymbol{B}}}_k$ are also lower-triangular.) By our assumption that $i \in \mathrm{pa}_{\mathcal{G}}(j) \Rightarrow i < j$, the node d in $\mathcal{G}$ has no child. Thus we can write

$$\boldsymbol{B}_k = \begin{pmatrix} \boldsymbol{B}_k^- & \boldsymbol{0} \\ \boldsymbol{b}_k & c_k \end{pmatrix}, \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}^- & \times \\ \times & \times \end{pmatrix} \text{ and } \hat{\hat{\boldsymbol{B}}}_k = \boldsymbol{B}_k \boldsymbol{T} = \begin{pmatrix} \hat{\hat{\boldsymbol{B}}}_k^- & \times \\ \times & \times \end{pmatrix}$$

where $\boldsymbol{B}_k^-, \boldsymbol{T}, \hat{\hat{\boldsymbol{B}}}_k^- = \boldsymbol{B}_k^- \boldsymbol{T}^- \in \mathbb{R}^{(d-1)\times(d-1)}, \boldsymbol{b}_k \in \mathbb{R}^{d-1}, c_k \in \mathbb{R}$ and $\times$ denotes irrelevant entries.

Let $\boldsymbol{A}_k^-, \hat{\boldsymbol{A}}_k^-, \boldsymbol{\Omega}_k^-$ and $\hat{\boldsymbol{\Omega}}_k^-$ be the top-left $(d-1)\times(d-1)$ sub-matrices of $\boldsymbol{A}_k, \hat{\boldsymbol{A}}, \boldsymbol{\Omega}_k$ and $\hat{\boldsymbol{\Omega}}_k$ respectively, and $\mathcal{G}^-$ and $\hat{\mathcal{G}}^-$ are graphs obtained by deleting node d and all related edges from $\mathcal{G}$ and $\hat{\mathcal{G}}$. Then it is easy to see that

$$\left(\boldsymbol{A}_k^-\right)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\mathcal{G}^-}(i) \quad \text{and} \quad \left(\hat{\boldsymbol{A}}_k^-\right)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\hat{\mathcal{G}}^-}(i). \tag{10}$$

Moreover,

$$\begin{pmatrix} \boldsymbol{B}_k^- & \boldsymbol{0} \\ \boldsymbol{b}_k & c_k \end{pmatrix} = \boldsymbol{B}_k = \boldsymbol{\Omega}_k^{-\frac{1}{2}} (\boldsymbol{I} - \boldsymbol{A}_k) = \begin{pmatrix} \left(\boldsymbol{\Omega}_k^-\right)^{-\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \times \end{pmatrix} \begin{pmatrix} \boldsymbol{I} - \boldsymbol{A}_k^- & \times \\ \times & \times \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{\Omega}_k^-\right)^{-\frac{1}{2}} (\boldsymbol{I} - \boldsymbol{A}_k^-) & \times \\ \times & \times \end{pmatrix}$$

so that $\boldsymbol{B}_k^- = \left(\boldsymbol{\Omega}_k^-\right)^{-\frac{1}{2}} (\boldsymbol{I} - \boldsymbol{A}_k^-)$. Similarly, we have $\hat{\hat{\boldsymbol{B}}}_k^- = \left(\hat{\boldsymbol{\Omega}}_k^-\right)^{-\frac{1}{2}} (\boldsymbol{I} - \hat{\boldsymbol{A}}_k^-)$.

We can also verify that $\left\{\boldsymbol{B}_k^-\right\}_{k=1}^K$ and $\left\{\hat{\hat{\boldsymbol{B}}}_k^-\right\}_{k=1}^K$ are node-level independent in the sense of Assumption 5. We only prove this for $\left\{\hat{\hat{\boldsymbol{B}}}_k\right\}_{k=1}^K$; the arguments used for $\{\boldsymbol{B}_k\}_{k=1}^K$ are exactly the same as the first case considered below. Now for each $i \in [d-1]$, let $\boldsymbol{R}_i \in \mathbb{R}^{K\times d}$ be the matrix whose k-th row is the i-th row of $\hat{\hat{\boldsymbol{B}}}_k$, and $\boldsymbol{R}_i^- \in \mathbb{R}^{K\times(d-1)}$ be the matrix whose k-th row is the i-th row of $\hat{\hat{\boldsymbol{B}}}_k^-$, then obviously $\boldsymbol{R}_i$ is of form $\left[\boldsymbol{R}_i^-, \boldsymbol{r}_i\right]$. We consider two cases:

- **Case 1.** $d \notin \mathrm{pa}_{\hat{\mathcal{G}}}(i)$ This means that the last entry of the i-th row of $\hat{\hat{\boldsymbol{B}}}_k$ is zero. Thus $\boldsymbol{r}_i = \boldsymbol{0}$, and $\mathrm{rank}\left(\boldsymbol{R}_i^-\right) = \mathrm{rank}\left(\boldsymbol{R}_i\right) = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(i)\right| = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}^-}(i)\right|$, where the second equality follows from Assumption 5.
- **Case 2.** $d \in \mathrm{pa}_{\hat{\mathcal{G}}}(i)$ In this case we have $\mathrm{rank}\left(\boldsymbol{R}_i^-\right) \geqslant \mathrm{rank}\left(\boldsymbol{R}_i\right) - 1 = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(i)\right| - 1 = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}^-}(i)\right|$. Due to our assumption on $\hat{\boldsymbol{A}}_k$ and the relationship $\hat{\hat{\boldsymbol{B}}}_k^- = \left(\hat{\boldsymbol{\Omega}}_k^-\right)^{-\frac{1}{2}} (\boldsymbol{I} - \hat{\boldsymbol{A}}_k^-)$, we know that each row of $\boldsymbol{R}_i^-$, namely the i-th row of some $\hat{\hat{\boldsymbol{B}}}_k$ only has $\left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(i)\right| - 1 = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}^-}(i)\right|$ non-zero entries, so that $\mathrm{rank}\left(\boldsymbol{R}_i^-\right) = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}^-}(i)\right|$ holds.

Since we have shown that the matrices $\boldsymbol{B}_k^-$ and $\hat{\hat{\boldsymbol{B}}}_k^-$ satisfy the three properties that we assume for induction with $\boldsymbol{T}$ replaced by $\boldsymbol{T}^-$ and $\mathcal{G}, \hat{\mathcal{G}}$ replaced by $\mathcal{G}^-, \hat{\mathcal{G}}^-$ respectively, by induction hypothesis, we can thus deduce that $\mathcal{G}^- = \hat{\mathcal{G}}^-$. To prove $\mathcal{G} = \hat{\mathcal{G}}$ it remains to show that the dependency of node d on the remaining nodes are the same in $\mathcal{G}$ and $\hat{\mathcal{G}}$.

First, we show that $\mathrm{ch}_{\hat{\mathcal{G}}}(d) = \emptyset$. Suppose in contrary that there is some $i \in \mathrm{ch}_{\hat{\mathcal{G}}}(d)$, then $\left|\mathrm{pa}_{\mathcal{G}}(i)\right| = \left|\mathrm{pa}_{\mathcal{G}^-}(i)\right| = \left|\mathrm{pa}_{\hat{\mathcal{G}}^-}(i)\right| = \left|\mathrm{pa}_{\hat{\mathcal{G}}}(i)\right| - 1$. Recalling that $(\boldsymbol{B})_i$ denotes the i-th row of matrix $\boldsymbol{B}$, we have

$$\begin{aligned} \dim\left(\mathrm{span}\left\langle\left(\hat{\hat{\boldsymbol{B}}}_k\right)_i : 1 \leqslant k \leqslant K\right\rangle\right) &= \dim\left(\mathrm{span}\left\langle(\boldsymbol{B}_k)_i : 1 \leqslant k \leqslant K\right\rangle\right) \\ &\leqslant \left|\mathrm{pa}_{\mathcal{G}}(i)\right| + 1 < \left|\mathrm{pa}_{\hat{\mathcal{G}}}(i)\right| + 1, \end{aligned} \tag{11}$$

where the first inequality follows from $\left(\hat{\hat{\boldsymbol{B}}}_k\right)_i = (\boldsymbol{B}_k)_i \boldsymbol{T}$ and Lemma 2, the second holds since each $(\boldsymbol{B}_k)_i$ has nonzero elements only at coordinates in $j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$, and the last one holds since

$\left|\mathrm{pa}_{\mathcal{G}}(i)\right| = \left|\mathrm{pa}_{\hat{\mathcal{G}}}(i)\right| - 1$. However, (11) contradicts the non-degeneracy condition Assumption 5 that we assume for matrices $\hat{\boldsymbol{B}}_k, k \in [K]$ in the statement of the theorem. Therefore we have $\mathrm{ch}_{\hat{\mathcal{G}}}(d) = \emptyset = \mathrm{ch}_{\mathcal{G}}(d)$.

Second, by a similar argument comparing the number of nonzero elements in the last row of $\boldsymbol{B}_k$ and $\hat{\hat{\boldsymbol{B}}}_k$, we can also deduce that

$$\left|\mathrm{pa}_{\mathcal{G}}(d)\right| = \left|\mathrm{pa}_{\hat{\mathcal{G}}}(d)\right|.$$

Indeed, since $\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d = (\boldsymbol{B}_k)_d \boldsymbol{T}$, by Lemma 2 we have

$$\dim\left(\mathrm{span}\left\langle\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d : 1 \leqslant k \leqslant K\right\rangle\right) = \dim\left(\mathrm{span}\left\langle(\boldsymbol{B}_k)_d : 1 \leqslant k \leqslant K\right\rangle\right)$$

However, since we assume that Assumption 5 is satisfied for $\{\boldsymbol{B}_k\}_{k=1}^K$ and $\left\{\hat{\boldsymbol{B}}_k\right\}_{k=1}^K$, we know that the LHS and RHS of the above equation are equal to $\left|\mathrm{pa}_{\mathcal{G}}(d)\right| + 1$ and $\left|\mathrm{pa}_{\hat{\mathcal{G}}}(d)\right| + 1$ respectively, implying (12).

Third, we show that $\mathrm{pa}_{\mathcal{G}}(d) = \mathrm{pa}_{\hat{\mathcal{G}}}(d)$. Suppose the contrary, let ℓ be the smallest element in $\mathrm{pa}_{\mathcal{G}}(d) \Delta \mathrm{pa}_{\hat{\mathcal{G}}}(d)$, where $A \Delta B := (A \setminus B) \cup (B \setminus A)$. Recall that while $\mathcal{G}$ and $\hat{\mathcal{G}}$ are originally not symmetric as nodes are topologically sorted according to $\mathcal{G}$, now we have shown that $\mathcal{G}^- \equiv \hat{\mathcal{G}}^-$ and that $\mathrm{ch}_{\mathcal{G}}(d) = \mathrm{ch}_{\hat{\mathcal{G}}}(d) = \emptyset$, so we can assume WLOG that $\ell \in \mathrm{pa}_{\mathcal{G}}(d)$ and $\ell \notin \mathrm{pa}_{\hat{\mathcal{G}}}(d)$, and the other case can be handled symmetrically. Since $\boldsymbol{B}_k$ is lower triangular and $(\boldsymbol{B}_k)_{jj} = (\Omega_k)_{jj}^{-\frac{1}{2}} \neq 0, \forall j \in [d]$, the top-left $\ell \times \ell$ sub-matrix of $\boldsymbol{B}_k$, which we denote by $[\boldsymbol{B}_k]_{\ell,\ell}$, must be invertible. This implies that $\left\{[\boldsymbol{B}_k]_{\ell,\ell}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathbb{R}^\ell\right\} = \mathbb{R}^\ell$, so we can always find coefficients $\lambda_{kj}, j \in [\ell]$ such that the first ℓ entries of the vector $(\boldsymbol{B}_k)_d - \sum_{i=1}^\ell \lambda_{ki}(\boldsymbol{B}_k)_i \in \mathbb{R}^d$ are all zero. Since $\hat{\hat{\boldsymbol{B}}}_k = \boldsymbol{B}_k \boldsymbol{T}$ and $\boldsymbol{T}$ is invertible, we have $\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d - \sum_{j=1}^\ell \lambda_{kj}\left(\hat{\hat{\boldsymbol{B}}}_k\right)_j = \left((\boldsymbol{B}_k)_d - \sum_{j=1}^\ell \lambda_{kj}(\boldsymbol{B}_k)_j\right)\boldsymbol{T}, \forall k \in [K]$ and

$$\begin{aligned}\dim\left(\mathrm{span}\left\langle\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d - \sum_{j=1}^\ell \lambda_{kj}\left(\hat{\hat{\boldsymbol{B}}}_k\right)_j : k \in [K]\right\rangle\right) &= \dim\left(\mathrm{span}\left\langle(\boldsymbol{B}_k)_d - \sum_{j=1}^\ell \lambda_{kj}(\boldsymbol{B}_k)_j : k \in [K]\right\rangle\right) \\ &\leqslant \left|\mathrm{pa}_{\mathcal{G}}(d) \setminus [\ell]\right| + 1.\end{aligned}$$

Here, the inequality holds because for any coordinate $t \in [d]$,

$$\left((\boldsymbol{B}_k)_d - \sum_{j=1}^\ell \lambda_{kj}(\boldsymbol{B}_k)_j\right)_t = \begin{cases} 0 & \text{if } t \leqslant \ell \\ (\boldsymbol{B}_k)_{d,t} & \text{otherwise} \end{cases} \tag{12}$$

where we note that $\boldsymbol{B}_k$ is lower-triangular and thus $(\boldsymbol{B}_k)_{j,t} = 0, \forall j \leqslant \ell, t > \ell$. This implies that $\left((\boldsymbol{B}_k)_d - \sum_{j=1}^\ell \lambda_{kj}(\boldsymbol{B}_k)_j\right)_t$ is nonzero only if $t > \ell$ and $t \in \mathrm{pa}_{\mathcal{G}}(d)$.

On the other hand, let $S = \left(\mathrm{pa}_{\hat{\mathcal{G}}}(d) \cap [\ell]^c\right) \cup \{d\}$, then

$$\begin{aligned}\dim\left(\mathrm{span}\left\langle\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d - \sum_{j=1}^\ell \lambda_{kj}\left(\hat{\hat{\boldsymbol{B}}}_k\right)_j : k \in [K]\right\rangle\right) &\geqslant \dim\left(\mathrm{span}\left\langle\left(\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d - \sum_{j=1}^\ell \lambda_{kj}\left(\hat{\hat{\boldsymbol{B}}}_k\right)_j\right)_S : k \in [K]\right\rangle\right) \\ &= \dim\left(\mathrm{span}\left\langle\left(\left(\hat{\hat{\boldsymbol{B}}}_k\right)_d\right)_S : k \in [K]\right\rangle\right) = |S|.\end{aligned}$$

where we recall that $\boldsymbol{u}_S$ denotes the vector $(u_i : i \in S) \in \mathbb{R}^{|S|}$. Here the first equality holds due to the same reason as (12), and the second follows from Assumption 5. To see why this is the case, note that Assumption 5 implies that the $K \times \left(\left|\mathrm{pa}_{\mathcal{G}}(d)\right| + 1\right)$ having $((\boldsymbol{B}_k)_d)_{\overline{\mathrm{pa}}_{\mathcal{G}}(d)}$ as the k-th row has full column rank, so that the sub-matrix obtained by extracting columns corresponding to the node set S also has full column rank.

We have shown that $\left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(d) \cap [\ell]^c\right| = |S| \leqslant \left|\mathrm{pa}_{\mathcal{G}}(d) \cap [\ell]^c\right| + 1 = \left|\overline{\mathrm{pa}}_{\mathcal{G}}(d) \cap [\ell]^c\right|$. On the other hand, recall that by our choice of ℓ, we have $\left|\overline{\mathrm{pa}}_{\mathcal{G}}(d) \cap [\ell-1]\right| = \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(d) \cap [\ell-1]\right|$ and $\ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(d) \setminus \overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(d)$. Putting these together, we have $\left|\overline{\mathrm{pa}}_{\mathcal{G}}(d)\right| > \left|\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(d)\right|$. However, we know from (12) that $\left|\mathrm{pa}_{\mathcal{G}}(d)\right| = \left|\mathrm{pa}_{\hat{\mathcal{G}}}(d)\right|$, leading to a contradiction. Hence, such ℓ shouldn't exist and we must have $\mathrm{pa}_{\mathcal{G}}(d) = \mathrm{pa}_{\hat{\mathcal{G}}}(d)$, completing the induction step for graphs of size d.

By the principle of induction, we have shown that $\mathcal{G} = \hat{\mathcal{G}}$ holds for any graphs under given assumptions. □

Now that we have established that $\mathcal{G} = \hat{\mathcal{G}}$, we prove the remaining part of the theorem. Note that for any $i, j \in [d]$ such that $i \notin \overline{\mathrm{pa}}_{\mathcal{G}}(j)$, we have $(\boldsymbol{B}_k)_{ji} = (\hat{\boldsymbol{B}}_k)_{ji} = 0, \forall k \in [K]$. Since $\hat{\boldsymbol{B}}_k = \boldsymbol{B}_k \boldsymbol{T}$, we have

$$\sum_{\ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(j)} (\boldsymbol{B}_k)_{j\ell} \boldsymbol{T}_{\ell i} = 0.$$

By Assumption 5, the above implies that $\boldsymbol{T}_{\ell i} = 0$ for $\forall \ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(j)$. In short, we have argued that if there exists j such that $i \notin \overline{\mathrm{pa}}_{\mathcal{G}}(j)$ and $\ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(j)$, then $\boldsymbol{T}_{\ell i} = 0$.

This implies that $T_{\ell i}$ is non-zero only if $\overline{\mathrm{ch}}_{\mathcal{G}}(\ell) \subseteq \overline{\mathrm{ch}}_{\mathcal{G}}(i)$. Since $\boldsymbol{v} = \boldsymbol{T}\boldsymbol{z}$, we have $\boldsymbol{v}_\ell = \sum_{i=1}^d \boldsymbol{T}_{\ell i} \boldsymbol{z}_i = \sum_{i \in [d]: \overline{\mathrm{ch}}_{\mathcal{G}}(\ell) \subseteq \overline{\mathrm{ch}}_{\mathcal{G}}(i)} \boldsymbol{T}_{\ell i} \boldsymbol{z}_i$. Note that when $i \neq \ell$, $\overline{\mathrm{ch}}_{\mathcal{G}}(\ell) \subseteq \overline{\mathrm{ch}}_{\mathcal{G}}(i)$ is equivalent to $i \in \mathrm{sur}_{\mathcal{G}}(\ell)$, so $\boldsymbol{v}_\ell$ only depends on $\boldsymbol{z}_{\overline{\mathrm{sur}}_{\mathcal{G}}(\ell)}$ by Lemma 6, as desired.

H.2 Formal version and proof of Theorem 2

In previous works [36, 54], it is common to consider single-node soft interventions in the following sense:

Assumption 8. *For $\forall 2 \leqslant k \leqslant K$, there exists $i_k \in [d]$, such that the structural equation in environment k satisfies (4) satisfies $\boldsymbol{w}_k(i) = \boldsymbol{w}_1(i)$ and $\omega_{k,i,i} = \omega_{1,i,i}$ for $\forall i \neq i_k$.*

Let $S_i = \{k : 2 \leqslant k \leqslant K, i_k = i\}, i \in [d]$ and $s_i = |S_i|$. Suppose that $\mathcal{G}$ has $e = \sum_{i=1}^d \left|\mathrm{pa}_{\mathcal{G}}(i)\right|$ edges, then we can view the weight vectors $\{(\boldsymbol{w}_k(i), \omega_{k,i,i}) : k = 1 \text{ or } i = i_k\}$ as elements of the Euclidean space $\mathbb{R}^{e+\sum_{k=2}^K \left|\mathrm{pa}_{\mathcal{G}}(i_k)\right|} \times \mathbb{R}_+^{d+K-1}$. Under Assumption 8, the models can be fully determined by these weight vectors. The following result states that if we restrict ourselves to single-node interventions, then in the worst case, $\Theta(d^2)$ interventions are required.

Theorem 6. *There exists a causal graph $\mathcal{G}$ with $\Theta(d^2)$ edges, such that for any unmixing matrix $\boldsymbol{H} \in \mathbb{R}^{d \times n}$ with full row rank, any independent noise variables ϵ, and any $s_i > 0, i \in [d]$ such that $s_i \leqslant \left|\mathrm{pa}_{\mathcal{G}}(i)\right|$ for some i, the following holds: except from a null set of the weight space $\mathbb{R}^{e+\sum_{k=2}^K \left|\mathrm{pa}_{\mathcal{G}}(i_k)\right|} \times \mathbb{R}_+^{d+K-1}$ (w.r.t the Lebesgue measure), there must exist a candidate solution $(\hat{\boldsymbol{H}}, \hat{\mathcal{G}})$ and a hypothetical data generating process*

$$\forall k \in [K], \quad \boldsymbol{v} = \hat{\boldsymbol{A}}_k \boldsymbol{v} + \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}} \epsilon, \quad \boldsymbol{x} = \hat{\boldsymbol{H}}^\dagger \boldsymbol{v}$$

such that

- *(i′) the unmixing matrix $\hat{\boldsymbol{H}} \in \mathbb{R}^{d \times n}$ has full row rank;*
- *(ii′) $\forall k \in [K]$ and $i, j \in [d]$, $(\hat{\boldsymbol{A}}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\hat{\mathcal{G}}}(i)$ and $\hat{\boldsymbol{\Omega}}_k$ is a diagonal matrix with positive entries;*
- *(iii′) for $\forall 2 \leqslant k \leqslant K$, the weight matrices $\hat{\boldsymbol{A}}_k, \hat{\boldsymbol{\Omega}}_k$ of environment E_k are from a single-node soft intervention on E_1 on node i_k, in the sense of Assumption 8,*

but $\mathcal{G}$ is non-isomorphic to $\hat{\mathcal{G}}$.

In this subsection we give the full proof of Theorem 6. We say that $S \subseteq \mathbb{R}^m$ is a null set if it has zero Lebesgue measure. Obviously, any hyperplanes in $\mathbb{R}^m$ are null sets. We will also need the following simple lemma:

Lemma 13. *Suppose that $m \in \mathbb{Z}_+$ and $\boldsymbol{V}$ is a subspace of $\mathbb{R}^m$. Then for any set of vectors $\boldsymbol{u}_i \in \mathbb{R}^m, i = 1, 2, \cdots, n$ that does not lie in $\boldsymbol{V}$, there must exists $\boldsymbol{v} \in \mathbb{R}^m$ such that $\boldsymbol{u}_i^\top \boldsymbol{v} \neq 0, \forall i \in [n]$ but $\boldsymbol{v} \in \boldsymbol{V}^\perp$, where $\boldsymbol{V}^\perp$ is the orthogonal space of $\boldsymbol{V}$.*

Proof. Let $\boldsymbol{w}_i$ be the orthogonal projection of $\boldsymbol{u}_i$ onto $\boldsymbol{V}^\perp$. Since $\boldsymbol{u}_i \notin \boldsymbol{V}$, we know that $\boldsymbol{w}_i \neq \boldsymbol{0}$. The solution space of each equation $\boldsymbol{w}_i^\top \boldsymbol{v} = 0$ in $\boldsymbol{V}^\perp$ must then be a proper subspace of $\boldsymbol{V}^\perp$. Equipped with the Lebesgue measure, all these spaces are null sets in $\boldsymbol{V}^\perp$, so one can always choose a $\boldsymbol{v} \in \boldsymbol{V}^\perp$ that does not lie in any of these solution spaces. Such $\boldsymbol{v}$ satisfies all the requirements. □

We choose $\mathcal{G}$ to be the graph with $i \to j$ for $\forall 1 \leqslant i < j \leqslant d$, so that $\mathcal{G}$ has $\frac{d(d-1)}{2}$ edges. Suppose that $i_0 \in [d]$ satisfies $s_i \leqslant \left|\mathrm{pa}_{\mathcal{G}}(i)\right| - 1$, then we must have $i_0 \geqslant 2$, so there is an edge $1 \to i_0$ in $\mathcal{G}$, Let $\hat{\mathcal{G}}$ be the resulting graph obtained via removing the edge $1 \to i_0$ in $\mathcal{G}$, then $\mathcal{G}$ and $\hat{\mathcal{G}}$ are clearly non-isomorphic.

Note that the i-th row of $\boldsymbol{B}_k$ can be written as $\omega_{k,i,i}^{-\frac{1}{2}} (\boldsymbol{e}_i - (\boldsymbol{A}_k)_i)$. Let's choose an lower-triangular matrix $\boldsymbol{T} = (t_{ij})_{i,j=1}^d \in \mathbb{R}^{d \times d}$ with columns $\boldsymbol{t}_i, i \in [d]$ such that the following holds:

$$(\boldsymbol{e}_i - (\boldsymbol{A}_k)_i)^\top \boldsymbol{t}_j = \begin{cases} = 0, & \forall k \in \{1\} \cup S_{i_0}, j = 1 \text{ and } i = i_0 \\ > 0, & \forall i = j \text{ and } k \in \{1\} \cup S_i \\ \neq 0, & \forall \text{ remaining } (i,j,k) \in \{k = 1, j < i\} \cup \{k \geqslant 2, i = i_k, j < i\} \end{cases} \tag{13}$$

and

$$t_{ii} \neq 0, \quad \forall i \in [d]. \tag{14}$$

We now show that: except from a null set in the weight space, such $\boldsymbol{T}$ can always be chosen. To see why this is the case, we first consider all the constraints on $\boldsymbol{t}_1$:

$$(\boldsymbol{e}_i - (\boldsymbol{A}_k)_i)^\top \boldsymbol{t}_1 = \begin{cases} = 0, & \forall k \in \{1\} \cup S_{i_0} \text{ and } i = i_0 \\ > 0, & \forall i = 1 \text{ and } k \in \{1\} \cup S_i \\ \neq 0, & \forall \text{ remaining } (i,k) \in \{k = 1, i > 1\} \cup \{k \geqslant 2, i = i_k > 1\} \end{cases} \tag{15}$$

Now let $\boldsymbol{V} = \mathrm{span}\,\langle \boldsymbol{e}_i - (\boldsymbol{A}_k)_i : k \in \{1\} \cup S_{i_0} \text{ and } i = i_0 \rangle$ and R be the set of pairs (i,k) specified in the second and third row of (15). For $\forall (i,k)$, let $\boldsymbol{w}_k(i)$ be the weight vector of node i in the environment k, *i.e.*, the vector of nonzero entries in $(\boldsymbol{A}_k)_i$. Then for $\forall (i,k) \in R$, the following set (as a subset of the weight space)

$$\bigcup_{k^* \in \{1\} \cup S_{i_0}} \{\boldsymbol{e}_{i_0} - \boldsymbol{w}_{k^*}(i_0) \in \mathrm{span}\,\langle \boldsymbol{e}_i - \boldsymbol{w}_k(i), \boldsymbol{e}_{i_0} - \boldsymbol{w}_{k'}(i_0) : k' \in \{1\} \cup S_{i_0} \setminus \{k^*\} \rangle\} \tag{16}$$

must be a null set. Thus

$$\boldsymbol{E} = \bigcup_{(i,k) \in R} \bigcup_{k^* \in \{1\} \cup S_{i_0}} \{\boldsymbol{e}_{i_0} - \boldsymbol{w}_{k^*}(i_0) \in \mathrm{span}\,\langle \boldsymbol{e}_i - \boldsymbol{w}_{k^*}(i), \boldsymbol{e}_{i_0} - \boldsymbol{w}_{k'}(i_0) : k' \in \{1\} \cup S_{i_0} \setminus \{k^*\} \rangle\} \tag{17}$$

is also a null set. For any weights that are not in $\boldsymbol{E}$, we necessarily have

$$\boldsymbol{e}_i - \boldsymbol{w}_k(i) \notin \mathrm{span}\,\langle \boldsymbol{e}_i - (\boldsymbol{A}_k)_i : k \in \{1\} \cup S_{i_0} \text{ and } i = i_0 \rangle = \boldsymbol{V}, \quad (i,k) \in R.$$

Let $U = \{\boldsymbol{e}_i - \boldsymbol{w}_k(i) : (i,k) \in R\}$, then we can apply Lemma 13 to deduce that there exists $\boldsymbol{t}_1$ such that

$$(\boldsymbol{e}_i - (\boldsymbol{A}_k)_i)^\top \boldsymbol{t}_1 = \begin{cases} = 0, & \forall k \in \{1\} \cup S_{i_0} \text{ and } i = i_0 \\ \neq 0, & \forall \text{ remaining } (i,k) \in \{k = 1\} \cup \{k \geqslant 2, i = i_k\} \end{cases} \tag{18}$$

Note that the only difference between (18) and (15) is that the latter one further requires that

$$(\boldsymbol{e}_1 - (\boldsymbol{A}_k)_1)^\top \boldsymbol{t}_1 > 0, \quad \forall k \in \{1\} \cup S_i.$$

while the former only guarantees that these terms are nonzero. However, recall that $(\boldsymbol{A}_k)_{ij} \neq 0 \Rightarrow j \in \mathrm{pa}_{\mathcal{G}}(i) \Rightarrow j < i$, so the above essentially says that $t_{11} > 0$. This can be easily guaranteed by replacing the solution $\boldsymbol{t}_1$ we obtained satisfying (18) with $-\boldsymbol{t}_1$ if needed.

Assuming that the weights do not lie in the null set $\boldsymbol{E}$ we have shown that $\boldsymbol{t}_1$ can always be chosen to satisfy all constraints imposed on it. We now proceed to choose the remaining entries of $\boldsymbol{T}$. The remaining entries in $\boldsymbol{t}_1$ can be chosen arbitrarily. For $\boldsymbol{t}_j, j > 1$, we note that the remaining constraints in (13) that need to be satisfied consist of the "nonzero" part and the "positivity" part. The positivity constrains can always be satisfied by choosing a sufficiently large t_{jj} for $j > 1$.

After choosing the $\boldsymbol{t}_j$'s satisfying the positivity constraints, the nonzero constraints along with (14) are easy to fulfill by slightly perturbing $\boldsymbol{t}_j$ if they are violated; since each of these constraints are only violated in a zero-measure set of the weight space. Hence, we have shown that except a null set $\boldsymbol{E}$ in the weight space, there always exists some $\boldsymbol{T}$ satisfying (13). Such $\boldsymbol{T}$ must be invertible since it is lower-triangular and its diagonal entries are nonzero. Now let $\hat{\boldsymbol{H}} = \boldsymbol{T}^{-1}\boldsymbol{H}$ and $\hat{\boldsymbol{\Omega}}_k$ be the diagonal matrix with entries $\hat{\omega}_{k,i,i} = t_{ii}^{-2} \cdot \omega_{k,i,i}, i \in [d]$ and

$$\hat{\boldsymbol{A}}_k = \boldsymbol{I} - \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}} \boldsymbol{\Omega}_k^{-\frac{1}{2}} (\boldsymbol{I} - \boldsymbol{A}_k)\boldsymbol{T}. \tag{19}$$

First since $\boldsymbol{T}$ is invertible and $\boldsymbol{H}$ has full row rank, $\hat{\boldsymbol{H}}$ must also have full row rank. Second,

$$(\hat{\boldsymbol{A}}_k)_{ij} = \begin{cases} 1 - \hat{\omega}_{k,i,i}^{\frac{1}{2}} \omega_{k,i,i}^{-\frac{1}{2}} t_{ii} = 0 & \text{if } j = i \\ -\hat{\omega}_{k,i,i}^{\frac{1}{2}} \omega_{k,i,i}^{-\frac{1}{2}} \left(\boldsymbol{e}_i - (\boldsymbol{A}_k)_i\right)^\top \boldsymbol{t}_j = 0 & \text{if } j > i \\ -\hat{\omega}_{k,i,i}^{\frac{1}{2}} \omega_{k,i,i}^{-\frac{1}{2}} \left(\boldsymbol{e}_i - (\boldsymbol{A}_k)_i\right)^\top \boldsymbol{t}_j & \text{if } j < i. \end{cases}$$

where we again recall that both $\boldsymbol{A}_k$ and $\boldsymbol{T}$ are lower-triangular. From (13) we can see that

- When $i = i_0$ and $j = 1$, we have
  - $(\hat{\boldsymbol{A}}_k)_{i_0,1} = 0$ if $k \in \{1\} \cup S_{i_0}$, and
  - $(\hat{\boldsymbol{A}}_k)_{i_0,1} = (\hat{\boldsymbol{A}}_1)_{i_0,1} = 0$ if $k \notin \{1\} \cup S_{i_0}$, by definition of S_{i_0} and Assumption 8.
- When $i > j$ and $(i,j) \neq (i_0, 1)$, we have
  - $(\hat{\boldsymbol{A}}_k)_{ij} \neq 0$ if $k = 1$ or $i = i_k$, which directly follows from (13), and
  - $(\hat{\boldsymbol{A}}_k)_{ij} = (\hat{\boldsymbol{A}}_1)_{ij} \neq 0$, by Assumption 8.

To summarize, for each k, $(\boldsymbol{A}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\mathcal{G}}(i)$ and $(i,j) \neq (i_0, 1)$.

Finally, let $\hat{\boldsymbol{w}}_k(i)$ be the weight vector of node i in environment k in the hypothetical model *i.e.*, the vector of nonzero entries in $(\boldsymbol{A}_k)_i$, and $\boldsymbol{T}_S$ be the submatrix of $\boldsymbol{T}$ by selecting the rows and columns in the index set S, then by (19) we have that

$$\hat{\omega}_{k,i,i} = t_{ii}^{-2} \cdot \omega_{k,i,i}, \quad \hat{\omega}_{k,i,i}^{\frac{1}{2}} \omega_{k,i,i}^{-\frac{1}{2}} \boldsymbol{w}_k(i) \boldsymbol{T}_{\mathrm{pa}_{\mathcal{G}}(i)} = \begin{cases} \hat{\boldsymbol{w}}_k(i) & \text{if } i \neq i_0 \\ [0, \hat{\boldsymbol{w}}_k(i)] & \text{if } i = i_0 \end{cases} \tag{20}$$

By our assumption, for $\forall k \geqslant 2$, $i \neq i_k \Rightarrow \boldsymbol{w}_k(i) = \boldsymbol{w}_1(i)$ and $\omega_{k,i,i} = \omega_{1,i,i}$. Thus (20) imply that $\forall k \geqslant 2$, $i \neq i_k \Rightarrow \hat{\boldsymbol{w}}_k(i) = \hat{\boldsymbol{w}}_1(i)$ and $\hat{\omega}_{k,i,i} = \hat{\omega}_{1,i,i}$. In other words, a single-node intervention on node i_k in environment k in the ground-truth model corresponds to a single-node intervention on node i_k in environment k in the hypothetical model, thereby completing the proof.

H.3 Proof of Theorem 4

We first prove two lemmas.

Lemma 14. *$\forall i \in [d]$, we have* $\mathrm{span}\left\langle (\boldsymbol{M}_k)_i : k \in [K] \right\rangle = \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \right\rangle$.

Proof. Since $(\boldsymbol{M}_k)_i = (\boldsymbol{B}_k)_i \boldsymbol{H}$, and $(\boldsymbol{B}_k)_{ij} \neq 0 \Leftrightarrow j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$, we can see that $(\boldsymbol{M}_k)_i \in \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \right\rangle$. On the other hand, since $\boldsymbol{H}$ is invertible, by Assumption 5 we have $\dim \mathrm{span}\left\langle (\boldsymbol{M}_k)_i : k \in [K] \right\rangle = \dim \mathrm{span}\left\langle (\boldsymbol{B}_k)_i : k \in [K] \right\rangle = \left|\overline{\mathrm{pa}}_{\mathcal{G}}(i)\right|$. Thus we must have $\mathrm{span}\left\langle (\boldsymbol{M}_k)_i : k \in [K] \right\rangle = \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \right\rangle$. □

Lemma 15. *Let $\hat{S}$ be an ancestral set of graph $\mathcal{G}$ and $\hat{\boldsymbol{V}}_k = \mathrm{span}\left\langle (\boldsymbol{M}_k)_s : s \in \hat{S} \right\rangle, k \in [K]$. Then we have $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \cdots = \boldsymbol{V}_K = \mathrm{span}\left\langle \boldsymbol{h}_s : s \in \hat{S} \right\rangle$.*

Proof. Recall that $\boldsymbol{M}_k = \boldsymbol{B}_k \boldsymbol{H}$, so for $\forall s \in \hat{S}$, the s-th row of $\boldsymbol{M}_k$ can be written as

$$(\boldsymbol{M}_k)_s = \sum_{t=1}^{d} (\boldsymbol{B}_k)_{st} \boldsymbol{h}_t = \sum_{t \in \overline{\mathrm{pa}}_{\mathcal{G}}(s)} (\boldsymbol{B}_k)_{st} \boldsymbol{h}_t \in \mathrm{span}\left\langle \boldsymbol{h}_s : s \in \hat{S} \right\rangle \tag{21}$$

where the last equation is because $\hat{S}$ is ancestral $\Rightarrow \overline{\mathrm{pa}}_{\mathcal{G}}(s) \subseteq \hat{S}$. Thus, for $\forall k \in [K]$, $\hat{\boldsymbol{V}}_k = \mathrm{span}\left\langle (\boldsymbol{M}_k)_s : s \in \hat{S} \right\rangle \subseteq \mathrm{span}\left\langle \boldsymbol{h}_s : s \in \hat{S} \right\rangle$. On the other hand, recall that both $\boldsymbol{B}_k$ and $\boldsymbol{H}$ have full rank, so $\boldsymbol{M}_k$ has full row rank as well, which implies that $\dim \boldsymbol{V}_k = |S| = \dim \mathrm{span}\left\langle \boldsymbol{h}_s : s \in \hat{S} \right\rangle$. Hence, $\boldsymbol{V}_k = \mathrm{span}\left\langle \boldsymbol{h}_s : s \in \hat{S} \right\rangle, \forall k \in [K]$. □

The following two propositions show that our algorithm always maintain an ancestral set, recursively adds a new node into the set and correctly identifies its parents.

Proposition 3 (Proposition 1 restated)**.** *The following two propositions hold for Algorithm 3:*

- $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S \Leftrightarrow$ *the* `if` *condition in line 8 of Algorithm 3 is fulfilled;*
- *the set S maintained in Algorithm 3 is always an ancestral set, in the sense that $j \in S \Rightarrow \mathrm{ans}_{\mathcal{G}}(j) \subseteq S$.*

Proof. At the starting point, we have $S = \emptyset$ which is obviously an ancestral set. Now suppose that after the ℓ-th iteration, $S = \{s_1, s_2, \cdots, s_\ell\}$ is an ancestral set. In the following, we show that $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S \Leftrightarrow$ the `if` condition in line 8 is fulfilled. This would immediately imply that there always exists a node i that can be added into S in the $(\ell+1)$-th iteration, and that after adding i, S is still an ancestral set.

Suppose that $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S$ for some $i \notin S$, by Lemma 14 we know that $(\boldsymbol{M}_k)_i \in \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \right\rangle$, so there exists $\alpha_k \in \mathbb{R}$ such that $(\boldsymbol{M}_k)_i - \alpha_k \boldsymbol{h}_i \in \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \mathrm{pa}_{\mathcal{G}}(i) \right\rangle$. Moreover, since $(\boldsymbol{M}_k)_i = \sum_{j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)} (\boldsymbol{B}_k)_{jj} \boldsymbol{h}_j$, $(\boldsymbol{B}_k)_{ii} = \omega_{k,i,i}^{-\frac{1}{2}} \neq 0$ and $\boldsymbol{H}$ has full row rank by assumption, we must have $(\boldsymbol{M}_k)_i \notin \mathrm{span}\left\langle \boldsymbol{h}_j : j \in \mathrm{pa}_{\mathcal{G}}(i) \right\rangle$ and so $\alpha_k \neq 0$. Thus, we have by the linearity of the projection operator

$$\boldsymbol{q}_k := \mathrm{proj}_{\boldsymbol{V}_k^\perp}\left((\boldsymbol{M}_k)_i\right) = \mathrm{proj}_{\boldsymbol{V}_k^\perp}\left((\boldsymbol{M}_k)_i - \alpha_k \boldsymbol{h}_i\right) + \mathrm{proj}_{\boldsymbol{V}_k^\perp}\left(\alpha_k \boldsymbol{h}_i\right) = \alpha_k \mathrm{proj}_{\boldsymbol{V}_k^\perp}\left(\boldsymbol{h}_i\right).$$

Recall that all the $\boldsymbol{V}_k$'s are the same and equal $\mathrm{span}\langle \boldsymbol{h}_s : s \in S \rangle$ by Lemma 15. So $\dim \mathrm{span}\langle \boldsymbol{q}_k : k \in [K] \rangle \leqslant 1$. Since $\boldsymbol{H}$ has full row rank, we have $\boldsymbol{h}_i \notin \mathrm{span}\langle \boldsymbol{h}_s : s \in S \rangle = \boldsymbol{V}_k$, so that $\dim \mathrm{span}\langle \boldsymbol{q}_k : k \in [K] \rangle = 1$ holds, which is exactly the `if` condition in line 8.

Conversely, suppose that there is an $i \notin S$ such that $\mathrm{ans}_{\mathcal{G}}(i) \nsubseteq S$ but $\dim \mathrm{span}\langle \boldsymbol{q}_k : k \in [K] \rangle = 1$ holds. Since S is ancestral, we know that there must be some $j \in \mathrm{pa}_{\mathcal{G}}(i)$ such that $j \notin S$. Since $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ both have support on the coordinates in $\overline{\mathrm{pa}}_{\mathcal{G}}(i)$, by Assumption 5 we know that $\mathrm{span}\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle \subseteq \mathrm{span}\langle (\boldsymbol{B}_k)_i : k \in [K] \rangle$, so that $\mathrm{span}\langle \boldsymbol{h}_i, \boldsymbol{h}_j \rangle = \mathrm{span}\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle \boldsymbol{H} \subseteq \mathrm{span}\langle (\boldsymbol{B}_k)_i : k \in [K] \rangle \boldsymbol{H} = \mathrm{span}\langle (\boldsymbol{M}_k)_i : k \in [K] \rangle$. Since $\dim \mathrm{span}\langle \boldsymbol{q}_k : k \in [K] \rangle = 1$, there must exist some vector $\boldsymbol{u} \in \mathbb{R}^n$ and $\alpha_i, \alpha_j \in \mathbb{R}$ such that $\boldsymbol{h}_i - \alpha_i \boldsymbol{u}, \boldsymbol{h}_j - \alpha_j \boldsymbol{u} \in \boldsymbol{V}_k = \mathrm{span}\langle \boldsymbol{h}_s : s \in S \rangle$. Since $i, j \notin S$ and $\boldsymbol{H}$ has full row rank, we can deduce that $\boldsymbol{h}_i, \boldsymbol{h}_j \notin \mathrm{span}\langle \boldsymbol{h}_s : s \in S \rangle$, and so both of α_i and α_j are non-zero. Hence $\alpha_j \boldsymbol{h}_i - \alpha_i \boldsymbol{h}_j \in \mathrm{span}\langle \boldsymbol{h}_s : s \in S \rangle$, which is impossible since we know that $\boldsymbol{H}$ has full row-rank. □

Proposition 4 (Proposition 2 restated)**.** *Given any ordered ancestral set S that contains $\mathrm{pa}_{\mathcal{G}}(i)$ for some $i \notin S$, Algorithm 2 returns a set $P_i \subseteq S$ that is exactly $\mathrm{pa}_{\mathcal{G}}(i)$.*

Proof. As we have shown in Proposition 1, for each possible input (S, i) to Algorithm 2, both S and $S \cup \{i\}$ are ancestral sets, so that $\mathrm{ans}_{\mathcal{G}}(i) \subseteq S$. Similarly one can see that inside the set $S := \{s_1, s_2, \cdots, s_m\}$, all the ancestors of s_j are contained in $\{s_1, s_2, \cdots, s_{j-1}\}$. In the following, we show that $\forall m' \in \{0, \ldots, m\}$, $r_{m'} = \left|\overline{\mathrm{pa}}_{\mathcal{G}}(i) - \{s_j : j \leqslant m'\}\right|$ (*).

By Lemma 15 we have $\boldsymbol{W}_1 = \boldsymbol{W}_2 = \cdots = \boldsymbol{W}_K = \operatorname{span}\left\langle \boldsymbol{h}_{s_j} : j \leqslant m' \right\rangle$. Let $t_1, t_2, \cdots, t_\ell$ be elements of $\overline{\mathrm{pa}}_{\mathcal{G}}(i)$ that are not in $\{s_j : j \leqslant m'\}$, then

$$
\begin{aligned}
r_{m'} &= \dim \operatorname{span}\left\langle \boldsymbol{p}_k : k \in [K] \right\rangle = \dim\left(\operatorname{proj}_{\operatorname{span}\left\langle \boldsymbol{h}_{s_j} : j \leqslant m' \right\rangle^{\perp}} \operatorname{span}\left\langle (\boldsymbol{M}_k)_i : k \in [K] \right\rangle\right) \\
&= \dim\left(\operatorname{proj}_{\operatorname{span}\left\langle \boldsymbol{h}_{s_j} : j \leqslant m' \right\rangle^{\perp}} \operatorname{span}\left\langle \boldsymbol{h}_j : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \right\rangle\right) \quad \text{(by Lemma 14)} \\
&= \dim\left(\operatorname{proj}_{\operatorname{span}\left\langle \boldsymbol{h}_{s_j} : j \leqslant m' \right\rangle^{\perp}} \operatorname{span}\left\langle \boldsymbol{h}_{t_1}, \boldsymbol{h}_{t_2}, \cdots, \boldsymbol{h}_{t_\ell} \right\rangle\right) \\
&= \ell \quad \text{(by Lemma 4 and non-degeneracy of } \boldsymbol{H}\text{)}
\end{aligned}
$$

which proves (*). From (*) it is easy to see that $m' \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$ (and thus in $\mathrm{pa}_{\mathcal{G}}(i)$ since $i \notin S$) if and only if $r_{m'} = r_{m'-1} - 1$. □

Now we conclude the proof of Theorem 4. Propositions 1 and 2 directly imply that Algorithm 3 is able to exactly recover the ground-truth causal graph $\mathcal{G}$. It remains to show that Line 20 in Algorithm 3 produces the correct $\hat{\boldsymbol{h}}_i$'s. By Lemma 14 we know that $E_j = \operatorname{span}\left\langle \boldsymbol{h}_\ell : \ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(j) \right\rangle$, so

$$
\cap_{j \in \overline{\mathrm{ch}}_{\mathcal{G}}(i)} E_j = \cap_{j \in \overline{\mathrm{ch}}_{\mathcal{G}}(i)} \operatorname{span}\left\langle \boldsymbol{h}_\ell : \ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(j) \right\rangle = \operatorname{span}\left\langle \boldsymbol{h}_\ell : \ell \in \overline{\mathrm{sur}}_{\mathcal{G}}(i) \right\rangle .
$$

where the last step holds because $\boldsymbol{H}$ has full row rank and $\cap_{j \in \overline{\mathrm{ch}}_{\mathcal{G}}(i)} \overline{\mathrm{pa}}_{\mathcal{G}}(j) = \overline{\mathrm{sur}}_{\mathcal{G}}(i)$ by definition. Hence, each $\hat{\boldsymbol{h}}_i$ is a linear combination of $\boldsymbol{h}_\ell, \ell \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, completing the proof.

I Identification limit of general causal models with soft interventions

While Theorem 1 guarantees identifiability with general environments, it only applies to linear causal models. In this section, we show that if we have access to single-node soft interventions, then we can identify general non-parametric causal models up to $\sim_{\mathrm{sur}}$. To obtain our identifiability result, we also require that the environments are non-degenerate in the following sense:

Definition 11 (Non-degeneracy set of interventions)**.** *Let $\hat{p}_k\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), k \in [K_i]$ be conditional probability densities at node i, then $\{\hat{p}_k\}_{k=1}^{K_i}$ is said to be non-degenerate on node i at point $\hat{\boldsymbol{z}} \in \mathbb{R}^d$ if all these conditional densities are well-defined and positive at $\hat{\boldsymbol{z}}$, and the matrix*

$$
\left[\frac{\partial\left(\hat{p}_1 / \hat{p}_k\right)}{\partial \boldsymbol{z}_j}\right]_{2 \leqslant k \leqslant K_i, j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)}\Bigg|_{\boldsymbol{z}=\hat{\boldsymbol{z}}} \in \mathbb{R}^{(K_i-1)\times\left(\left|\mathrm{pa}_{\mathcal{G}}(i)\right|+1\right)}
$$

has full row rank. Moreover, we say that $\{\hat{p}_k\}_{k=1}^{K_i}$ is non-degenerate in a point set $\boldsymbol{O}$ if for all $\hat{\boldsymbol{z}} \in \boldsymbol{O}$, it is non-degenrate at $\hat{\boldsymbol{z}}$.

The following lemma shows how Definition 11 is related to Assumption 5 in the linear setting:

Lemma 16. *Suppose that $\hat{p}_k(\boldsymbol{z}) = \prod_{i=1}^d \hat{p}_k\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), k \in [K]$ be probability distributions of latent variables $\boldsymbol{z}$ generated from the linear causal models (3), such that for $\forall i \in [d]$, $\hat{p}_k\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), k \in [K]$ are non-degenerate on node i in the sense of Definition 11. Then the corresponding matrices $\boldsymbol{B}_k, k \in [K]$ satisfy Assumption 5.*

Now we are ready to state our main result in this section:

Theorem 7. *Suppose that we have access to observations generated from multiple environments $\{P_{\boldsymbol{X}}^E\}_{E \in \mathfrak{E}}$. Let $\left(\hat{\boldsymbol{h}}, \hat{\mathcal{G}}\right)$ be any candidate solution with data generated according to Assumption 1 with latent variables $\boldsymbol{v} = \hat{\boldsymbol{h}}(\boldsymbol{x})$ and joint distribution q_E with factors q_i^E. Assuming that*

- *(i) the joint densities $\{p_E(\boldsymbol{z})\}_{E \in \mathfrak{E}}$ are continuous differentiable on $\mathbb{R}^d$ with common support $\boldsymbol{O}_{\boldsymbol{z}}$, and $\{q_E(\boldsymbol{v})\}_{E \in \mathfrak{E}}$ are continuous differentiable on $\mathbb{R}^d$ with common support $\boldsymbol{O}_{\boldsymbol{v}}$;*
- *(ii) we have access to multiple single-node soft interventions on each node with unknown targets: there exists a partition $\mathfrak{E} = \cup_{i=1}^d \mathfrak{E}_i$ such that $\mathcal{I}_{\boldsymbol{z}}^{\mathfrak{E}_i} = \{\pi(i)\}, \mathcal{I}_{\boldsymbol{v}}^{\mathfrak{E}_i} = \{\pi'(i)\}, \forall i \in [d]$ for some unknown permutations π and π' on $[d]$;*

(iii) the intervention distributions on each node are non-degenerate in the sense of Definition 11: there exists $\boldsymbol{N_z} \subseteq \boldsymbol{O_z}$ and $\boldsymbol{N_v} \subseteq \boldsymbol{O_v}$ satisfying $\boldsymbol{N_z^o} = \boldsymbol{N_v^o} = \emptyset$ where S^o denotes the interior of a set S, such that for all $i \in [d]$, $\left\{p_i^E(\cdot) : E \in \mathfrak{E}_{\pi^{-1}(i)}\right\}$ (resp. $\left\{q_i^E(\cdot) : E \in \mathfrak{E}_{\pi'^{-1}(i)}\right\}$) is non-degenerate on node i in $\boldsymbol{O_z} \setminus \boldsymbol{N_z}$ (resp. $\boldsymbol{O_v} \setminus \boldsymbol{N_v}$).

Then we must have $(\boldsymbol{h}, \mathcal{G}) \sim_{\text{sur}} (\hat{\boldsymbol{h}}, \hat{\mathcal{G}})$.

Previous works on the identifiability of non-parametric causal models typically require that all the joint distributions are supported on the whole space $\mathbb{R}^d$ [49, 23, 47]. In contrast, we only assume that the densities have common and unknown support across all interventions.

Theorem 7 can be regarded as a soft-intervention version of 49, Theorem 4.3, which assumes access to hard interventions and only need two paired interventions per node. While they are able to show full identifiability, we show in the following that identifiability up to $\sim_{\text{sur}}$ is the best we can hope for with soft interventions.

Theorem 8 (Counterpart to Theorem 7, informal version of Theorem 10)**.** *For any causal model $(\boldsymbol{h}, \mathcal{G})$ and any set of environments $\mathfrak{E} = \{E_k : k \in [K]\}$ such that all conditions in Theorem 7 are satisfied, there must exists a candidate solution $(\hat{\boldsymbol{h}}, \mathcal{G})$ and a hypothetical data generating process that satisfy the same set of conditions, but*

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad \forall j \in \overline{\text{sur}}_{\mathcal{G}}(i).$$

Finally, the ambiguity still exists if we additionally assume standard axioms such as causal minimality (Assumption 6) and faithfulness (Assumption 7) on the causal model.

I.1 Proof of Lemma 16

Let $\boldsymbol{w}_k(i) \in \mathbb{R}^{\left|\text{pa}_{\mathcal{G}}(i)\right|}$ be the vector obtained by removing all zero entries in the i-th row of $\boldsymbol{A}_k$ and $\omega_{k,i,i}$ be the i-th diagonal entry in $\boldsymbol{\Omega}_k$, then for the k-th environment we have $\boldsymbol{z}_i = \boldsymbol{w}_k(i)^\top \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)} + \omega_{k,i,i}^{\frac{1}{2}} \epsilon_i$, so that

$$\hat{p}_k\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\right) = \omega_{k,i,i}^{-\frac{1}{2}} p_{\epsilon_i}\left(\omega_{k,i,i}^{-\frac{1}{2}}(\boldsymbol{z}_i - \langle \boldsymbol{w}_k(i), \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\rangle)\right)$$

where $p_{\epsilon_i}(\cdot)$ is the density of ϵ_i. As a result, we have

$$\begin{aligned}
\nabla \frac{\hat{p}_1}{\hat{p}_k}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\right) &= \frac{\hat{p}_1}{\hat{p}_k}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\right) \cdot \nabla \log \frac{\hat{p}_1}{\hat{p}_k}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\right) \\
&= \frac{\hat{p}_1}{\hat{p}_k}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\right) \cdot [c_{i1}(1, -\boldsymbol{w}_1(i)) - c_{ik}(1, -\boldsymbol{w}_k(i))]
\end{aligned}$$

where for convenience we use ∇ to denote the gradient with respect to all variables $\boldsymbol{z}_{\overline{\text{pa}}_{\mathcal{G}}(i)}$, and $c_{ik} = \omega_{k,i,i}^{-\frac{1}{2}} \cdot \frac{p'_{\epsilon_i}}{p_{\epsilon_i}}\left(\omega_{k,i,i}^{-\frac{1}{2}}(\boldsymbol{z}_i - \langle \boldsymbol{w}_k(i), \boldsymbol{z}_{\text{pa}_{\mathcal{G}}(i)}\rangle\right)$ (we omit the dependency on $\boldsymbol{z}$ for simplicity).

Definition 11 implies that $\text{span}\left\langle c_{i1}(1, -\boldsymbol{w}_1(i)) - c_{ik}(1, -\boldsymbol{w}_k(i)) : 2 \leqslant k \leqslant K\right\rangle = \mathbb{R}^{\left|\text{pa}_{\mathcal{G}}(i)\right|+1}$, thus it holds that $\text{span}\langle(1, -\boldsymbol{w}_k(i)) : k \in [K]\rangle = \mathbb{R}^{\left|\text{pa}_{\mathcal{G}}(i)\right|+1}$ as well. By definition of $\boldsymbol{B}_k$, this immediately implies that $\dim\left(\text{span}\left\langle(\boldsymbol{B}_k)_i : k \in [K]\right\rangle\right) = \left|\text{pa}_{\mathcal{G}}(i)\right| + 1$ as desired.

I.2 Proof of Theorem 7

Define $\boldsymbol{\tau} := \hat{\boldsymbol{h}} \circ \boldsymbol{h}^{-1} : \mathbb{R}^d \mapsto \mathbb{R}^d$, then we have that $\boldsymbol{v} = \boldsymbol{\tau}(\boldsymbol{z})$. Since both $\boldsymbol{h}$ and $\hat{\boldsymbol{h}}$ are diffeomorphisms by assumption, so is $\boldsymbol{\tau}$. To avoid confusion, in this section we use $\boldsymbol{z}$ (resp. $\boldsymbol{v}$) to denote random variables while using $\hat{\boldsymbol{z}}$ (resp. $\hat{\boldsymbol{v}}$) to denote (deterministic) vectors.

Let $\mathfrak{E}_j = \left\{E_k^{(j)} : k \in [K_j]\right\}$ be the j-th collection of environments according to our assumption. We first prove the following lemma:

Lemma 17. $\boldsymbol{O_v} = \boldsymbol{\tau}(\boldsymbol{O_z})$.

Proof. By the change of variable formula [35], for $\forall \hat{\boldsymbol{z}} \in \mathbb{R}^d$ and $\forall E \in \mathfrak{E}$ we have $p_E(\hat{\boldsymbol{z}}) = q_E(\hat{\boldsymbol{v}}) \left|\det \boldsymbol{J}_{\boldsymbol{\tau}}(\hat{\boldsymbol{z}})\right|$, where $\hat{\boldsymbol{v}} = \boldsymbol{\tau}(\hat{\boldsymbol{z}})$. Since $\boldsymbol{\tau}$ is a diffeomorphism, we must have $\left|\det \boldsymbol{J}_{\boldsymbol{\tau}}(\hat{\boldsymbol{z}})\right| \neq 0$, so $\hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}} \Leftrightarrow \hat{\boldsymbol{v}} = \boldsymbol{\tau}(\hat{\boldsymbol{z}}) \in \boldsymbol{O}_{\boldsymbol{v}}$, concluding the proof. □

Lemma 18. *Let $\hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}$. For $\forall j \in [d]$ and $2 \leqslant k \leqslant K_j$, we have*

$$\frac{p_j^{E_k^{(j)}}}{p_j^{E_1^{(j)}}}\left(\hat{\boldsymbol{z}}_j \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) = \frac{q_j^{E_k^{(j)}}}{q_j^{E_1^{(j)}}}\left(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(j)}\right), \tag{22}$$

where $\hat{\boldsymbol{v}} = \boldsymbol{\tau}(\hat{\boldsymbol{z}}) \in \boldsymbol{O}_{\boldsymbol{v}}$.

Proof. Since $\boldsymbol{v} = \boldsymbol{\tau}(\boldsymbol{z})$, by the change-of-measure formula [35] we have that for $\forall \hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}$,

$$\prod_{i=1}^{d} p_i^E\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) = p_E(\hat{\boldsymbol{z}}) = q_E(\hat{\boldsymbol{v}}) \left|\det \boldsymbol{J}_{\boldsymbol{\tau}}(\hat{\boldsymbol{z}})\right| = \prod_{i=1}^{d} q_i^E\left(\boldsymbol{\tau}_i(\hat{\boldsymbol{z}}) \mid \boldsymbol{\tau}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}(\hat{\boldsymbol{z}})\right) \left|\det \boldsymbol{J}_{\boldsymbol{\tau}}(\hat{\boldsymbol{z}})\right| \tag{23}$$

for all $E \in \mathfrak{E}_j$, where $\hat{\boldsymbol{v}} = \boldsymbol{\tau}(\hat{\boldsymbol{z}})$. By Assumption (ii) and Definition 2, we know that $p_i^{E_k^1} = p_i^{E_1^{(1)}} \Leftrightarrow i \neq 1$ and $q_i^{E_k^1} = q_i^{E_1^{(1)}} \Leftrightarrow i \neq 1$ for all $k > 1$. Thus, we have that

$$\prod_{i=1}^{d} \frac{p_i^{E_k^{(j)}}\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)}{p_i^{E_1^{(j)}}\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)} = \frac{p_j^{E_k^{(j)}}}{p_j^{E_1^{(j)}}}(\hat{\boldsymbol{z}}_j \mid \hat{\boldsymbol{z}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)})$$

and

$$\prod_{i=1}^{d} \frac{q_i^{E_k^{(j)}}\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}\right)}{q_i^{E_1^{(j)}}\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}\right)} = \frac{q_j^{E_k^{(j)}}}{q_j^{E_1^{(j)}}}(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}).$$

Since the LHS of the above two equations are the same by (23), the RHS must also be the same, concluding the proof. □

We assume WLOG that the vertices of $\mathcal{G}$ are labelled such that $i \to j \Rightarrow i < j$, and that $\pi(i) = i, \forall i \in [d]$. Also we can assume the nodes are fixed and only consider how they are connected, *i.e.,* $\pi'(i) = i, \forall i \in [d]$. [1]

Lemma 19. *We have $(\boldsymbol{\tau}(\boldsymbol{N}_{\boldsymbol{z}}))^{\mathrm{o}} = \left(\boldsymbol{\tau}^{-1}(\boldsymbol{N}_{\boldsymbol{v}})\right)^{\mathrm{o}} = \emptyset$.*

Proof. The result immediately follows from the assumption that $\boldsymbol{N}_{\boldsymbol{z}}^{\mathrm{o}} = \boldsymbol{N}_{\boldsymbol{v}}^{\mathrm{o}} = \emptyset$ and that τ is a diffeomorphism. □

For any vertex set V, we use $\mathcal{G}_V$ to denote its corresponding induced subgraph of $\mathcal{G}$. We first prove the following statements by induction on j:

(1) $\forall i \neq j, i \in \mathrm{pa}_{\mathcal{G}}(j) \Leftrightarrow i \in \mathrm{pa}_{\mathcal{G}'}(j)$;

(2) $\forall j \in [d]$, there exists a continuously differentiable function ϕ_i such that $\boldsymbol{v}_j = \phi_j\left(\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}\right)$. Moreover, $\frac{\partial \phi_j}{\partial \boldsymbol{z}_j} \not\equiv 0$ (*i.e.*, not always zero).

(3) $\forall j \in [d]$, there exists a continuously differentiable function Υ_j such that $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)} = \Upsilon_j(\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)})$.

For $j = 1$, by assumption $\mathrm{pa}_{\mathcal{G}}(j) = \emptyset$. Lemma 18 implies that for any $\hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}$,

$$\frac{p_1^{E_k^{(1)}}}{p_1^{E_1^{(1)}}}(\hat{\boldsymbol{z}}_1) = \frac{q_1^{E_k^{(1)}}}{q_1^{E_1^{(1)}}}\left(\hat{\boldsymbol{v}}_1 \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(1)}\right), \forall 2 \leqslant k \leqslant K_1. \tag{24}$$

[1] This is also WLOG because we now have groups of soft interventions where each group corresponds to a single node, so we can just relabel the node in $\hat{\mathcal{G}}$ that corresponds to the i-th group as node i.

Then for $\forall i \in \overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(1)$, taking the partial derivative w.r.t $\boldsymbol{v}_j$ gives

$$\frac{\partial}{\partial \hat{\boldsymbol{v}}_i} \frac{q_1^{E_k^{(1)}}}{q_1^{E_1^{(1)}}} \left(\hat{\boldsymbol{v}}_1 \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(1)}\right) = \begin{pmatrix} p_1^{E_k^{(1)}} \\ p_1^{E_1^{(1)}} \end{pmatrix}' (\hat{\boldsymbol{z}}_1) \cdot \frac{\partial \hat{\boldsymbol{z}}_1}{\partial \hat{\boldsymbol{v}}_i} \Rightarrow \nabla_{\boldsymbol{v}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(1)}} \frac{q_1^{E_k^{(1)}}}{q_1^{E_1^{(1)}}} \left(\hat{\boldsymbol{v}}_1 \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(1)}\right) = \begin{pmatrix} p_1^{E_k^{(1)}} \\ p_1^{E_1^{(1)}} \end{pmatrix}' (\hat{\boldsymbol{z}}_1) \cdot \nabla_{\boldsymbol{v}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(1)}} \hat{\boldsymbol{z}}_1.$$

Thus,

$$\operatorname{rank}\left[\nabla_{\boldsymbol{v}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(1)}} \frac{q_1^{E_k^{(1)}}}{q_1^{E_1^{(1)}}} \left(\hat{\boldsymbol{v}}_1 \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(1)}\right) : 2 \leqslant k \leqslant K_1\right] \leqslant 1.$$

Note that the above inequality holds for $\forall \hat{\boldsymbol{v}} \in \boldsymbol{O}_{\boldsymbol{v}}$. If $\mathrm{pa}_{\hat{\mathcal{G}}}(1) \neq \emptyset$, then this would contradict the non-degeneracy assumption *(iii)* which implies that the above matrix should have rank $\geqslant 2$ at some point $\hat{\boldsymbol{v}} \in \boldsymbol{O}_{\boldsymbol{v}}$. Hence we must have $\mathrm{pa}_{\hat{\mathcal{G}}}(1) = \emptyset$, implying that (1) holds for $j = 1$.

Taking the derivative of both sides of (24) w.r.t $\boldsymbol{z}_i, i \geqslant 2$ implies that $\begin{pmatrix} q_1^{E_k^{(1)}} \\ q_1^{E_1^{(1)}} \end{pmatrix}' (\hat{\boldsymbol{v}}_1) \cdot \frac{\partial \hat{\boldsymbol{v}}_1}{\partial \hat{\boldsymbol{z}}_i} = 0$. By our assumption *(iii)*, for $\forall \hat{\boldsymbol{v}} \in \boldsymbol{O}_{\boldsymbol{v}} \setminus \boldsymbol{N}_{\boldsymbol{v}}$, there exists $2 \leqslant k \leqslant K_1$ such that $\begin{pmatrix} q_1^{E_k^{(1)}} \\ q_1^{E_1^{(1)}} \end{pmatrix}' (\hat{\boldsymbol{v}}_1) \neq 0$, and thus we have $\frac{\partial \hat{\boldsymbol{v}}_1}{\partial \hat{\boldsymbol{z}}_i} = 0, \forall \hat{\boldsymbol{z}} \in \boldsymbol{\tau}^{-1}(\boldsymbol{O}_{\boldsymbol{v}} \setminus \boldsymbol{N}_{\boldsymbol{v}})$. Since $\boldsymbol{\tau}$ is a diffeomorphism, we can deduce that $\boldsymbol{\tau}^{-1}(\boldsymbol{O}_{\boldsymbol{v}} \setminus \boldsymbol{N}_{\boldsymbol{v}}) = \boldsymbol{O}_{\boldsymbol{z}} \setminus \boldsymbol{\tau}^{-1}(\boldsymbol{N}_{\boldsymbol{v}})$ and $\left(\boldsymbol{\tau}^{-1}(\boldsymbol{N}_{\boldsymbol{v}})\right)^{\mathrm{o}} = \emptyset$ by Lemma 19. As a result, we actually have $\frac{\partial \hat{\boldsymbol{v}}_1}{\partial \hat{\boldsymbol{z}}_i} = 0, \forall \hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}$. Hence in $\boldsymbol{O}_{\boldsymbol{z}}$ there exists a continuous differentiable function ϕ_1 such that $\boldsymbol{v}_1 = \phi_1(\boldsymbol{z}_1)$, proving (2). Finally, (3) directly follows from (2) since $\mathrm{pa}_{\mathcal{G}}(1) = \emptyset$, concluding the proof for $j = 1$.

Now suppose that the statement holds up to $j-1$, and we need to prove it for j. Again by Lemma 18 we have for $\forall \hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}$ that

$$\frac{p_j^{E_k^{(j)}}}{p_j^{E_1^{(j)}}} \left(\hat{\boldsymbol{z}}_j \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) = \frac{q_j^{E_k^{(j)}}}{q_j^{E_1^{(j)}}} \left(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(j)}\right), \quad \forall 2 \leqslant k \leqslant K_j. \tag{25}$$

For all $i \notin \overline{\mathrm{pa}}_{\mathcal{G}}(j)$, taking partial derivative w.r.t. $\boldsymbol{z}_i$ gives

$$0 = \sum_{\ell \in \overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)} \frac{\partial}{\partial \hat{\boldsymbol{v}}_\ell} \frac{q_j^{E_k^{(j)}}}{q_j^{E_1^{(j)}}} \left(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(j)}\right) \cdot \frac{\partial \hat{\boldsymbol{v}}_\ell}{\partial \hat{\boldsymbol{z}}_i}, \quad \forall 2 \leqslant k \leqslant K_j,$$

i.e.,

$$\left[\nabla_{\boldsymbol{v}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}} \frac{q_j^{E_k^{(j)}}}{q_j^{E_1^{(j)}}} \left(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\hat{\mathcal{G}}}(j)}\right) : 2 \leqslant k \leqslant K_j\right]^\top \frac{\partial \hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}}{\partial \hat{\boldsymbol{z}}_i} = 0.$$

Similar to the $j = 1$ case, by assumption *(iii)*, we know that the above corfficient matrix has full row rank for $\forall \hat{\boldsymbol{v}} \in \boldsymbol{O}_{\boldsymbol{v}} \setminus \boldsymbol{N}_{\boldsymbol{v}}$, so for $\forall \boldsymbol{z} \in \boldsymbol{\tau}^{-1}(\boldsymbol{O}_{\boldsymbol{v}} \setminus \boldsymbol{N}_{\boldsymbol{v}}) = \boldsymbol{O}_{\boldsymbol{z}} \setminus \boldsymbol{\tau}^{-1}(\boldsymbol{N}_{\boldsymbol{v}})$, we have $\frac{\partial \hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}}{\partial \hat{\boldsymbol{z}}_i} = 0$. Since $\left(\boldsymbol{\tau}^{-1}(\boldsymbol{N}_{\boldsymbol{v}})\right)^{\mathrm{o}} = \emptyset$ by Lemma 19, for all $\hat{\boldsymbol{z}} \in \boldsymbol{N}_{\boldsymbol{z}}$ we can choose a sequence of points $\hat{\boldsymbol{z}}^{(i)}, i = 1, 2, \cdots$ in $\boldsymbol{O}_{\boldsymbol{z}}$ such that $\hat{\boldsymbol{z}}^{(i)} \to \hat{\boldsymbol{z}}$. Since $\boldsymbol{\tau}$ is a diffeomorphism, its derivatives are continuous and we can deduce that $\frac{\partial \hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}}{\partial \hat{\boldsymbol{z}}_i} = \lim_{\ell \to +\infty} \frac{\partial \hat{\boldsymbol{v}}^{(\ell)}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}}{\partial \hat{\boldsymbol{z}}_i^{(\ell)}} = 0$. As a result, $\frac{\partial \hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)}}{\partial \hat{\boldsymbol{z}}_i} = 0$ actually holds for all $\boldsymbol{z} \in \boldsymbol{O}_{\boldsymbol{z}}$. Hence, there exists a continuous differentiable function Υ_j such that $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)} = \Upsilon_j\left(\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}\right)$.

By our assumption, $\mathrm{pa}_{\mathcal{G}}(j) \subseteq [j-1]$. Suppose that $\mathrm{pa}_{\hat{\mathcal{G}}}(j) \nsubseteq \{i : i < j\}$, let $\ell \in \mathrm{pa}_{\hat{\mathcal{G}}}(j) \setminus \{i : i < j\}$, then by induction hypothesis, $\hat{\boldsymbol{v}}_t = \boldsymbol{\tau}_t(\hat{\boldsymbol{z}}), \hat{\boldsymbol{z}} \in \boldsymbol{O}_{\boldsymbol{z}}, t = 1, 2, \cdots, j, \ell$ are all functions of $\hat{\boldsymbol{z}}_1, \cdots, \hat{\boldsymbol{z}}_j$. Since $\boldsymbol{\tau}$ is a diffeomorphism and $\boldsymbol{O}_{\boldsymbol{z}}$ is the support of the distributions $p_E, E \in \mathfrak{E}$, we can deduce that the support of the latent variables $(\boldsymbol{v}_t : t = 1, 2, \cdots, j, \ell)$ lie on a submanifold with dimension $\leqslant j$, which is impossible since $\boldsymbol{v}$ is supported on the open set $\boldsymbol{O}_{\boldsymbol{v}} \subseteq \mathbb{R}^d$ by assumption *(i)*.

Hence, we must have $\mathrm{pa}_{\hat{\mathcal{G}}}(j) \subseteq \{i : i < j\}$. Furthermore, if there exists $i \in \mathrm{pa}_{\hat{\mathcal{G}}}(j)$ such that $i \notin \mathrm{pa}_{\mathcal{G}}(j)$, then the induction hypothesis implies that $\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_i} \not\equiv 0$, but $\boldsymbol{v}_i$ is a function of $\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}$ as previously derived, which is also a contradiction. Thus we actually have $\mathrm{pa}_{\hat{\mathcal{G}}}(j) \subseteq \mathrm{pa}_{\mathcal{G}}(j)$.

In a completely symmetric manner, we can take the derivatives of (25) w.r.t. $\boldsymbol{v}_i, \forall i \in \overline{\mathrm{pa}}_{\hat{\mathcal{G}}}(j)$ and obtain that $\mathrm{pa}_{\mathcal{G}}(j) \subseteq \mathrm{pa}_{\hat{\mathcal{G}}}(j)$. Hence, $\mathrm{pa}_{\hat{\mathcal{G}}}(j) = \mathrm{pa}_{\mathcal{G}}(j)$, completing the proof of (1) and (3) for the j case.

Finally, if $\frac{\partial \boldsymbol{v}_j}{\partial \boldsymbol{z}_j} \equiv 0$, then by (3) and the induction hypothesis, $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_j$ are all functions of $\boldsymbol{z}_{[j-1]}$, which implies that $(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_j)$ lies on a submanifold with dimension $\leqslant j-1$, again contradicting assumption *(i)*. Thus $\frac{\partial \boldsymbol{v}_j}{\partial \boldsymbol{z}_j} \not\equiv 0$. This completes the proof of our inductive step.

To recap, we now know that

- $\mathcal{G} = \hat{\mathcal{G}}$, and
- For $\forall i \in [d]$, there exists a function Υ_i such that $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)} = \Upsilon_i\left(\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}\right)$.

It remains to show that for $\forall k \in \mathrm{pa}_{\mathcal{G}}(i) \setminus \mathrm{sur}_{\mathcal{G}}(i)$, Υ_i doesn't depend on $\boldsymbol{z}_k$.

By definition, if $k \in \mathrm{pa}_{\mathcal{G}}(i) \setminus \mathrm{sur}_{\mathcal{G}}(i)$, we know that there exists $j \in \mathrm{ch}_{\mathcal{G}}(i)$ such that $j \notin \mathrm{ch}_{\mathcal{G}}(k)$. We have shown that $\boldsymbol{v}_i$, as a component of $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}$, is a function of $\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(j)}$. By the choice of k, we have $k \notin \overline{\mathrm{pa}}_{\mathcal{G}}(j)$, so that $\boldsymbol{v}_i$ does not depend on $\boldsymbol{z}_k$. The conclusion follows.

J Omitted Proofs for Theorem 3 and Theorem 8

In this section we provide detailed proofs of main ambiguity results.

Definition 12. *We say that a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is effect-respecting for a causal graph $\mathcal{G}$, or $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}(\mathcal{G})$, if $\boldsymbol{M}_{ij} \neq 0 \Leftrightarrow j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$. We also write $\boldsymbol{M} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$ if $\boldsymbol{M}$ is invertible and $\boldsymbol{M}_{ij} \neq 0 \Rightarrow j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$. Finally, we write $\boldsymbol{M} \in \overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ if $\boldsymbol{M}_{ij} \neq 0 \Rightarrow j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$.*

Remark 1. *By definition $\mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$ is the set of all matrices $\boldsymbol{M}$ where $\boldsymbol{M}_{ij} \neq 0, \forall j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i)$, so it can be identified as $\mathbb{R}^{d+d_{\mathcal{G}}}$ where $d_{\mathcal{G}} = \sum_{i=1}^{d} |\mathrm{sur}_{\mathcal{G}}(i)|$. Equipped with the Lebesgue measure, we have $\mathcal{M}_{\mathrm{sur}}(\mathcal{G}) \subset \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G}) \subset \overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ and $\overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G}) \setminus \mathcal{M}_{\mathrm{sur}}(\mathcal{G})$ is a null set. In the remaining part of this section, we will use measure-theoretic statement for $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}(\mathcal{G})$ in the above sense.*

We first present a result that serves as a good starting point to understand why this is the case. It states that latent representations that are equivalent under $\sim_{\mathrm{sur}}$ are essentially generated from the same causal graph.

Proposition 5. *Let $\boldsymbol{M}$ be an invertible matrix such that $\boldsymbol{M}_{ij} \neq 0 \Rightarrow j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i)$. Suppose that the latent variables $\boldsymbol{z} \in \mathbb{R}^d$ are generated from any distributions $p_i\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), i \in [d]$ with joint density $p(\boldsymbol{z}) = \prod_{i=1}^{d} p_i\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)$, then the joint density of $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$ can be written as $q(\boldsymbol{v}) = \prod_{i=1}^{d} q_i\left(\boldsymbol{v}_i \mid \boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)$ for some density functions $q_i, i \in [d]$.*

J.1 Proof of Proposition 5

We first prove the following lemma:

Lemma 20. *Let $\boldsymbol{M} \in \mathcal{M}^0_{\mathrm{sur}}(\mathcal{G})$ and latent variables $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$, then for $\forall i \in [d]$, there exists invertible matrices $\boldsymbol{M}_i$ and $\boldsymbol{M}_i^-$ such that $\boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)} = \boldsymbol{M}_i^- \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}$ and $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)} = \boldsymbol{M}_i \boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$.*

Proof. $\forall j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$, we know that $\boldsymbol{v}_j$ is a linear function of $\boldsymbol{z}_\ell, \ell \in \overline{\mathrm{sur}}_{\mathcal{G}}(j)$. By Lemma 7, we know that $\overline{\mathrm{sur}}_{\mathcal{G}}(j) \subseteq \overline{\mathrm{pa}}_{\mathcal{G}}(i)$, so each $\boldsymbol{v}_j, j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)$ is a linear function of $\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$. Thus we can write $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)} = \boldsymbol{M}_i \boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$. In the following we argue that $\boldsymbol{M}_i$ is invertible. Let π be a permutation on $\overline{\mathrm{pa}}_{\mathcal{G}}(i)$ such that $k \in \mathrm{pa}_{\mathcal{G}}(\ell) \Rightarrow \pi(k) < \pi(\ell)$ (such π can always be chosen since $\mathcal{G}$ is acyclic), then we can write

$$\left(\hat{\boldsymbol{v}}_{\pi(j)} : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)\right)^\top = \tilde{\boldsymbol{M}}_i \left(\hat{\boldsymbol{z}}_{\pi(j)} : j \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)\right)^\top \tag{26}$$

where $\tilde{\boldsymbol{M}}_i$ is an upper triangular matrix with non-zero diagonal entries by our choice of $\boldsymbol{M}$. Since $\boldsymbol{M}_i$ can be obtained from $\tilde{\boldsymbol{M}}_i$ be exchanging a few rows and columns, $\boldsymbol{M}_i$ is invertible as well.

Similarly, using the fact that $\forall j \in \mathrm{pa}_{\mathcal{G}}(i)$, $\overline{\mathrm{sur}}_{\mathcal{G}}(j) \subseteq \mathrm{pa}_{\mathcal{G}}(i)$, we can prove the existence of an invertible matrix $\boldsymbol{M}_i^-$ such that $\boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)} = \boldsymbol{M}_i^- \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}$. □

Returning to the proof of Proposition 5. Assume WLOG that the nodes of $\mathcal{G}$ are ordered in a way such that $i \in \mathrm{pa}_{\mathcal{G}}(j) \Rightarrow i < j$, so that $\boldsymbol{M}$ is a lower-triangular matrix. The joint density of $\boldsymbol{v}$ can be written as

$$q(\boldsymbol{v}) = \prod_{i=1}^{d} q\left(\boldsymbol{v}_i \mid \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{i-1}\right).$$

Since $\boldsymbol{v} = \boldsymbol{M}z$ and $\boldsymbol{M}$ is lower triangular and invertible (hence, with non-zero diagonals), we know that $(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{i-1})$ is an invertible linear function of $(\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{i-1})$ and $(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_i)$ is an invertible linear function of $(\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_i)$. Let $\hat{\boldsymbol{v}} = \hat{\boldsymbol{M}}\hat{\boldsymbol{z}} \in \mathbb{R}^d$, then we have

$$\begin{aligned} q\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_1, \cdots, \hat{\boldsymbol{v}}_{i-1}\right) &= \frac{q(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \cdots, \hat{\boldsymbol{v}}_i)}{q(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \cdots, \hat{\boldsymbol{v}}_{i-1})} = \frac{p(\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_i) \det \hat{\boldsymbol{M}}_{1:i,1:i}}{p(\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_{i-1}) \det \hat{\boldsymbol{M}}_{1:i-1,1:i-1}} \\ &\propto \frac{p(\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_i)}{p(\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_{i-1})} = p\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_1, \cdots, \hat{\boldsymbol{z}}_{i-1}\right) = p_i\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), \end{aligned}$$

where $\hat{\boldsymbol{M}}_{1:i,i:i}$ denotes that top-left submatrix of $\hat{\boldsymbol{M}}$ of size $i \times i$, and the last step follows from the causal Markov condition (Definition 1). On the other hand, let $q_i\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)$ be the conditional density of $\boldsymbol{v}_i$ on its parents at $\hat{\boldsymbol{v}} \in \mathbb{R}^d$. For $\forall j \in \mathrm{pa}_{\mathcal{G}}(i)$, from $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$ we know that $\boldsymbol{v}_j$ is a linear function of $\boldsymbol{z}_{\overline{\mathrm{sur}}_{\mathcal{G}}(j)}$. By Lemma 20 we know that $\hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}$ is a linear function of $\hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}$ and $\hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$ is a linear function of $\hat{\boldsymbol{z}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$, so that

$$q\left(\hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \propto p\left(\hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \quad \text{and} \quad q\left(\hat{\boldsymbol{v}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}\right) \propto p\left(\hat{\boldsymbol{z}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}\right)$$

and

$$q_i\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \propto \frac{p\left(\hat{\boldsymbol{z}}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}\right)}{p\left(\hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right)} = p_i\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right).$$

Hence, we have $q_i\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \propto q\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_1, \cdots, \hat{\boldsymbol{v}}_{i-1}\right)$, so that

$$q(\hat{\boldsymbol{v}}) = \prod_{i=1}^{d} q_i\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \propto \prod_{i=1}^{d} q_i\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right).$$

Since both sides integrate to 1, it turns out that they are equal, as desired.

J.2 Formal version and proof of Theorem 3: the linear case

Theorem 9 (Counterpart to Theorem 1)**.** *For any causal model $(\boldsymbol{H}, \mathcal{G})$ and any set of environments $\mathfrak{E} = \{E_k : k \in [K]\}$, suppose that we have observations $\left\{P_{\boldsymbol{X}}^E\right\}_{E \in \mathfrak{E}}$ satisfying Assumption 1:*

$$\forall k \in [K], \quad \boldsymbol{z} = \boldsymbol{A}_k \boldsymbol{z} + \boldsymbol{\Omega}_k^{\frac{1}{2}} \epsilon, \quad \boldsymbol{x} = \boldsymbol{H}^{\dagger} \boldsymbol{z}$$

such that

- *(i) the unmixing matrix $\boldsymbol{H} \in \mathbb{R}^{d \times n}$ has full row rank;*
- *(ii) $\forall k \in [K]$ and $i, j \in [d]$, $(\boldsymbol{A}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\mathcal{G}}(i)$ and $\boldsymbol{\Omega}_k$ is a diagonal matrix with positive entries;*
- *(iii) $\left\{\boldsymbol{B}_k = \boldsymbol{\Omega}_k^{-\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{A}_k)\right\}_{k=1}^{K}$ are node level non-degenerate in the sense of Assumption 5,*

then there must exist a candidate solution $(\hat{\boldsymbol{H}}, \mathcal{G})$ and a hypothetical data generating process

$$\forall k \in [K], \quad \boldsymbol{v} = \hat{\boldsymbol{A}}_k \boldsymbol{v} + \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}} \epsilon, \quad \boldsymbol{x} = \hat{\boldsymbol{H}}^{\dagger} \boldsymbol{v}$$

such that

(i′) *the unmixing matrix* $\hat{\boldsymbol{H}} \in \mathbb{R}^{d\times n}$ *has full row rank;*

(ii′) $\forall k \in [K]$ *and* $i, j \in [d]$, $(\hat{\boldsymbol{A}}_k)_{ij} \neq 0 \Leftrightarrow j \in \mathrm{pa}_{\mathcal{G}}(i)$ *and* $\hat{\boldsymbol{\Omega}}_k$ *is a diagonal matrix with positive entries;*

(iii′) $\left\{\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{\Omega}}_k^{-\frac{1}{2}}(\boldsymbol{I} - \hat{\boldsymbol{A}}_k)\right\}_{k=1}^{K}$ *are node level non-degenerate in the sense of Assumption 5,*

but

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad \forall j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i).$$

Finally, if we additionally assume that

(iii) *the environments are groups of single-node interventions: there exists a partition* $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$ *such that* $\mathcal{I}_{\boldsymbol{z}}^{\mathfrak{E}_i} = \{i\}$ *(see Definition 2),*

then we can guarantee the existence of $(\hat{\boldsymbol{H}}, \mathcal{G})$ *and weight matrices which, besides the properties listed above, also satisfy*

(iii′) *for the same partition* $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$, *we have* $\mathcal{I}_{\boldsymbol{v}}^{\mathfrak{E}_i} = \{i\}$.

In other words, additionally assuming that the environments are from single-node interventions does not resolve the ambiguity.

Remark 2. *Compared with our identifiability guarantee Theorem 1, Theorem 9 actually demonstrates a stronger form of impossibility. Specifically, it states that the SNA cannot be resolved even if both the ground-truth causal graph and the noise variables are known.*

We define

$$\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z} \tag{27}$$

where $\boldsymbol{M}$ is an effect-respecting matrix. At this point we do not make any other restrictions on $\boldsymbol{M}$, but we will specify the appropriate choise of $\boldsymbol{M}$ later.

By assumption, the latent variables in the k-th environment are generated by

$$\boldsymbol{z} = \boldsymbol{A}_k\boldsymbol{z} + \boldsymbol{\Omega}_k^{\frac{1}{2}}\epsilon,$$

then $\boldsymbol{v} = \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{A}_k)^{-1}\boldsymbol{\Omega}_k^{\frac{1}{2}}\epsilon$. Let $\hat{\boldsymbol{\Omega}}_k$ be the diagonal matrix with entries $\boldsymbol{M}_{ii}^2 \cdot (\boldsymbol{\Omega}_k)_{ii}, i \in [d]$ and $\hat{\boldsymbol{A}}_k = \boldsymbol{I} - \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}}\boldsymbol{\Omega}_k^{-\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{A}_k)\boldsymbol{M}^{-1}$, then $\boldsymbol{v} = \hat{\boldsymbol{A}}_k\boldsymbol{v} + \hat{\boldsymbol{\Omega}}_k^{\frac{1}{2}}\epsilon$. Note that the choice of $\hat{\boldsymbol{\Omega}}_k$ here is to that the diagonal entries of $\hat{\boldsymbol{A}}_k$ are zero, as we show below. It remains to show that: for almost all $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}^0(\mathcal{G})$, it holds for $\forall k \in [K]$ that $(\hat{\boldsymbol{A}}_k)_{ij} = 0 \Leftrightarrow j \notin \mathrm{pa}_{\mathcal{G}}(i)$.

For the $\Leftarrow$ direction, since $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}^0(\mathcal{G})$, $\boldsymbol{M}^{-1} \in \mathcal{M}_{\mathrm{sur}}^0(\mathcal{G})$ as well. Thus, $\forall j \notin \mathrm{pa}_{\mathcal{G}}(i)$ we have

$$\begin{aligned}\left[(\boldsymbol{I} - \boldsymbol{A}_k)\boldsymbol{M}^{-1}\right]_{ij} &= \sum_{\ell=1}^{d}(\boldsymbol{I} - \boldsymbol{A}_k)_{i\ell} \cdot (\boldsymbol{M}^{-1})_{\ell j} = \sum_{\ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(i) \cap \{\ell' : j \in \overline{\mathrm{sur}}_{\mathcal{G}}(\ell')\}} (\boldsymbol{I} - \boldsymbol{A}_k)_{i\ell} \cdot (\boldsymbol{M}^{-1})_{\ell j} \\ &= \begin{cases} 0 & \text{if } j \notin \overline{\mathrm{pa}}_{\mathcal{G}}(i) \\ (\boldsymbol{M}^{-1})_{ii} & \text{if } j = i \end{cases}\end{aligned}$$

where the last step holds because $\forall \ell \in [d]$, $\ell \in \overline{\mathrm{pa}}_{\mathcal{G}}(i), j \in \overline{\mathrm{sur}}_{\mathcal{G}}(\ell) \Rightarrow j \in \overline{\mathrm{pa}}_i$, and when $j = i$, the only such ℓ is $\ell = i$. Hence, we can see that our choice of $\hat{\boldsymbol{A}}_k$ satisfies

$$\left(\hat{\boldsymbol{A}}_k\right)_{ij} = \begin{cases} 0 - 0 = 0 & \text{if } j \notin \overline{\mathrm{pa}}_{\mathcal{G}}(i) \\ 1 - \hat{\omega}_{k,i,i}^{\frac{1}{2}}\omega_{k,i,i}^{-\frac{1}{2}}(\boldsymbol{M}^{-1})_{ii} = 0 & \text{if } j = i, \end{cases}$$

so $\left(\hat{\boldsymbol{A}}_k\right)_{ij} \neq 0 \Rightarrow j \in \mathrm{pa}_{\mathcal{G}}(i)$.

Conversely, for $\forall j \in \mathrm{pa}_{\mathcal{G}}(i)$,

$$(\hat{\boldsymbol{A}}_k)_{ij} = 0 \Leftrightarrow \sum_{s \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)} (\boldsymbol{I} - \boldsymbol{A}_k)_{is}(\boldsymbol{M}^{-1})_{sj} = 0 \Leftrightarrow \sum_{s \in \overline{\mathrm{pa}}_{\mathcal{G}}(i)} (-1)^s (\boldsymbol{I} - \boldsymbol{A}_k)_{is} \det \boldsymbol{M}_{sj}^{-} = 0 \tag{28}$$

where $\boldsymbol{M}_{sj}^{-}$ is the $(d-1)\times(d-1)$ matrix obtained by removing the s-th row and j-th column of $\boldsymbol{M}$, and the second step in the equation above follows from the fact that $\boldsymbol{M}^{-1} = \det(\boldsymbol{M})^{-1}\mathrm{adj}(\boldsymbol{M})$, where $\mathrm{adj}(\boldsymbol{M})$ denotes the adjugate matrix of $\boldsymbol{M}$ whose (i,j)-th entry is $(-1)^{i+j}\det \boldsymbol{M}_{ij}^{-}$.

(28) holds if only if $\boldsymbol{M}$ takes values on a lower-dimensional algebraic manifold of its embedded space $\mathbb{R}^{d+d_{\mathcal{G}}}$ (see Remark 1). As a result, for almost every $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}^{0}(\mathcal{G})$, $\boldsymbol{v}$ is generated from a linear causal model with graph $\mathcal{G}$ as defined in (3). Moreover, let $\hat{\boldsymbol{B}}_k = \boldsymbol{B}_k\boldsymbol{M}^{-1}, k \in [K]$, so that $\epsilon = \hat{\boldsymbol{B}}_k\boldsymbol{v}$ in the k-th environment. Then for all nodes $i \in [d]$ and $S \subseteq \mathrm{pa}(i)\cup\{i\}$, we have

$$\begin{aligned}\dim \mathrm{span}\left\langle \left(\hat{\boldsymbol{B}}_k{}^{\top}\boldsymbol{e}_i\right)_S : k\in[K]\right\rangle &= \dim \mathrm{span}\left\langle \boldsymbol{M}^{-\top}\left(\left(\boldsymbol{B}_k^{\top}\boldsymbol{e}_i\right)_S : k\in[K]\right)\right\rangle \\ &= \dim \mathrm{span}\left\langle \left(\boldsymbol{B}_k^{\top}\boldsymbol{e}_i\right)_S : k\in[K]\right\rangle = \left|\mathrm{pa}_{\mathcal{G}}(i)\right| + 1,\end{aligned}$$

implying that $\hat{\boldsymbol{B}}_k, k\in[K]$ satisfy Assumption 5.

Now we have shown that for almost every $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}^{0}(\mathcal{G})$, we can construct a hypothetical data generating process with latent variables $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$ that satisfies all requirements in Theorem 9. Choose an arbitrary $\boldsymbol{M}$ that is in $\mathcal{M}_{\mathrm{sur}}(\mathcal{G})$, then we have that

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad j \notin \overline{\mathrm{sur}}_{\mathcal{G}}(i).$$

Finally, if we additionally assume single-node interventions, $\forall k,\ell \in \mathfrak{E}_i$, we have that $(\boldsymbol{B}_k)_j \neq (\boldsymbol{B}_\ell)_j \Leftrightarrow j = i$. For any $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}^{0}(\mathcal{G})$ (and specifically the $\boldsymbol{M}$ that we have already chosen above), we have $(\hat{\boldsymbol{B}}_k)_j = (\boldsymbol{B}_k)_j\boldsymbol{M}^{-1}$ and $(\hat{\boldsymbol{B}}_\ell)_j = (\boldsymbol{B}_\ell)_j\boldsymbol{M}^{-1}, \forall j\in[d]$. Thus, $(\hat{\boldsymbol{B}}_k)_j \neq (\hat{\boldsymbol{B}}_\ell)_j \Leftrightarrow j = i$ as well, implying that $\mathfrak{E}_i$ is also a group of single-node interventions on $\boldsymbol{v}$, concluding the proof.

J.3 Formal statement and proof of Theorem 10: the non-parametric case

Theorem 10 (Counterpart to Theorem 7)**.** *For any causal model $(\boldsymbol{h},\mathcal{G})$ and any set of environments $\mathfrak{E}$, suppose that we have observations $\left\{P_{\boldsymbol{X}}^{E}\right\}_{E\in\mathfrak{E}}$ satisfying Assumption 1:*

$$\forall E\in\mathfrak{E}, \boldsymbol{z}\sim p_E(\hat{\boldsymbol{z}}) = \prod_{i=1}^{d} p_i^E\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), \boldsymbol{x} = \boldsymbol{h}^{-1}(\boldsymbol{z})$$

such that

- *(i) all densities p_i^E are continuously differentiable and the joint density p_E is positive everywhere;*
- *(ii) the environments are groups of single-node interventions: there exists a partition $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$ such that $\mathcal{I}_{\boldsymbol{z}}^{\mathfrak{E}_i} = \{i\}$;*
- *(iii) the intervention distributions on each node are non-degenerate: $\forall i\in[d]$, the set of distributions $\left\{p_i^E : E\in\mathfrak{E}_i\right\}$ satisfy Definition 11 at any point $\hat{\boldsymbol{z}}\in\mathbb{R}^d$,*

then there must exist a candidate solution $(\hat{\boldsymbol{h}},\mathcal{G})$ and a hypothetical data generating process

$$\forall E\in\mathfrak{E}, \boldsymbol{v}\sim q_E(\hat{\boldsymbol{v}}) = \prod_{i=1}^{d} q_i^E\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_{\mathrm{pa}_{\mathcal{G}}(i)}\right), \boldsymbol{x} = \hat{\boldsymbol{h}}^{-1}(\boldsymbol{v})$$

such that

- *(i′) all densities q_i^E are continuously differentiable and the joint density q_E is positive everywhere;*
- *(ii′) for the same partition $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$, we have $\mathcal{I}_{\boldsymbol{v}}^{\mathfrak{E}_i} = \{i\}$;*
- *(iii′) $\forall i\in[d]$, the set of distributions $\left\{q_i^E : E\in\mathfrak{E}_i\right\}$ satisfy Definition 11 at any point $\hat{\boldsymbol{v}}\in\mathbb{R}^d$,*

but

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad \forall j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i).$$

Remark 3. *Similar to the case of Theorem 9, Appendix J.3 also establishes a stronger form of identifiability. First, it is assumed that the causal graph $\mathcal{G}$ is known. Second, we only focus on a special case of the setting of Theorem 7 by assuming that the support is the whole space, and the non-degeneracy condition Definition 11 holds at any point. Even in this case, we show that our identification guarantee up to SNA cannot be improved.*

We state and prove a stronger version of Theorem 10:

Theorem 11. *For any causal model $(\boldsymbol{h}, \mathcal{G})$ and any set of environments $\mathfrak{E}$, suppose that we have observations $\left\{P_{\boldsymbol{X}}^E\right\}_{E\in\mathfrak{E}}$ satisfying Assumption 1:*

$$\forall E \in \mathfrak{E}, \quad \boldsymbol{z} \sim p_E(z) = \prod_{i=1}^{d} p_i^E\left(z_i \mid z_{\mathrm{pa}_{\mathcal{G}}(i)}\right), \quad \boldsymbol{x} = \boldsymbol{h}^{-1}(\boldsymbol{z})$$

such that

- *(i) all densities p_i^E are continuously differentiable and the joint density p_E is positive everywhere;*
- *(ii) the environments are groups of single-node interventions: there exists a partition $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$ such that $\mathcal{I}_{\boldsymbol{z}}^{\mathfrak{E}_i} = \{i\}$;*
- *(iii) the intervention distributions on each node are non-degenerate: $\forall i \in [d]$, the set of distributions $\left\{p_i^E : E \in \mathfrak{E}_i\right\}$ satisfy Definition 11,*

then there must exist a candidate solution $(\hat{\boldsymbol{h}}, \mathcal{G})$ and a hypothetical data generating process

$$\forall E \in \mathfrak{E}, \quad \boldsymbol{v} \sim q_E(v) = \prod_{i=1}^{d} q_i^E\left(v_i \mid v_{\mathrm{pa}_{\mathcal{G}}(i)}\right), \quad \boldsymbol{x} = \hat{\boldsymbol{h}}^{-1}(\boldsymbol{v})$$

such that

- *(i′) all densities q_i^E are continuously differentiable and the joint density q_E is positive everywhere;*
- *(ii′) for the same partition $\mathfrak{E} = \cup_{i=1}^{d}\mathfrak{E}_i$, we have $\mathcal{I}_{\boldsymbol{v}}^{\mathfrak{E}_i} = \{i\}$;*
- *(iii′) $\forall i \in [d]$, the set of distributions $\left\{q_i^E : E \in \mathfrak{E}_i\right\}$ satisfy Definition 11,*

but

$$\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j} \neq 0, \quad \forall j \in \overline{\mathrm{sur}}_{\mathcal{G}}(i).$$

Finally, if we additionally assume minimality (Assumption 6) and/or faithfulness (Assumption 7) of all p_E's, we can guarantee the existence of $(\hat{\boldsymbol{h}}, \mathcal{G})$ and q_E's satisfying minimality and/or faithfulness in addition to the properties listed above. In other words, assuming minimality and/or faithfulness does not resolve the ambiguity.

Proof. We define

$$\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z} \tag{29}$$

where $\boldsymbol{M}$ is an effect-respecting matrix. At this point we do not make any other restrictions on $\boldsymbol{M}$, and we will choose appropriate $\boldsymbol{M}$ later. By Lemma 20, there exists invertible matrices $\boldsymbol{M}_i$ and $\boldsymbol{M}_i^-$ such that $\boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)} = \boldsymbol{M}_i^- \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}$ and $\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)} = \boldsymbol{M}_i \boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}$, so for all environment $E \in \mathfrak{E}$ we have

$$q_i^E(\boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)}) = p_i^E(\boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot \left|\det(\boldsymbol{M}_i^-)^{-1}\right|, \quad q_i^E(\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}) = p_i^E(\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}) \cdot \left|\det(\boldsymbol{M}_i)^{-1}\right|$$

so that

$$q_i^E\left(\boldsymbol{v}_i \mid \boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) = p_i^E\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) \frac{\left|\det \boldsymbol{M}_i^{-1}\right|}{\left|\det(\boldsymbol{M}_i^-)^{-1}\right|}, \quad \forall i \in [d]. \tag{30}$$

In the following, assuming that $\left(p_i^E : E \in \mathfrak{E}\right)$ satisfies any of the listed assumptions, we show that $\left(q_i^E : E \in \mathfrak{E}\right)$ satisfies the same assumption as well.

Firstly, (30) immediately implies that the density of $\boldsymbol{v}$ is continuous differentiable and positive everywhere. Secondly, $\forall k, \ell \in \mathfrak{E}_i$, we have that

$$p_j^{E_k}\left(\boldsymbol{z}_j \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) = p_j^{E_\ell}\left(\boldsymbol{z}_j \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) \Leftrightarrow j = i.$$

By (30) it is easy to see that

$$q_j^{E_k}\left(\boldsymbol{v}_j \mid \boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) = q_j^{E_\ell}\left(\boldsymbol{v}_j \mid \boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(j)}\right) \Leftrightarrow j = i$$

as well, *i.e.,* $q^k, k \in \mathfrak{E}_i$ are single-node interventions on $\boldsymbol{v}_i$ according to Definition 2.

Thirdly, we verify the non-degeneracy condition for q_i^E's. Indeed we have for $\forall k \geqslant 2$ that

$$\nabla_{\boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}} \frac{q_i^{E_1}}{q_i^{E_k}}\left(\boldsymbol{v}_i \mid \boldsymbol{v}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) = \frac{\partial \boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}}{\partial \boldsymbol{v}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}} \nabla_{\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}} \frac{q_i^{E_1}}{q_i^{E_k}}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right) = \boldsymbol{M}_i^{-1} \nabla_{\boldsymbol{z}_{\overline{\mathrm{pa}}_{\mathcal{G}}(i)}} \frac{q_i^{E_1}}{q_i^{E_k}}\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}}(i)}\right).$$

Since $\boldsymbol{M}_i$ is invertible, the above equation and the non-degeneracy of $p^{E_k}, k \in [K]$ immediately implies that non-degeneracy of $q^{E_k}, k \in [K]$.

Thus, for arbitrary $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}(\mathcal{G})$, we have constructed a hypothetical data generating process with latent variable $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$ that satisfies all given conditions. It remains to show that such construction is still possible under additional minimality and faithfulness conditions.

Claim 1. There exists a neighbourhood O of the identity matrix $\boldsymbol{I}$ in $\overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ (in the sense of Remark 1) such that for $\forall \boldsymbol{M} \in O \cap \mathcal{M}_{\mathrm{sur}}^0(\mathcal{G})$, $p^{E_k}, k \in [K]$ satisfy Assumption 7 $\Rightarrow q^{E_k}, k \in [K]$ satisfy Assumption 7.

For $\forall i, j$ not d-separated by $S \subseteq [d]$, for all $k \in [K]$ there exists $\hat{\boldsymbol{z}} \in \mathbb{R}^d$ such that $\Delta_k^{(i,j,S)} = p^{E_k}\left(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{z}}_j \mid \hat{\boldsymbol{z}}_S\right) - p^{E_k}\left(\hat{\boldsymbol{z}}_i \mid \hat{\boldsymbol{z}}_S\right) p^{E_k}\left(\hat{\boldsymbol{z}}_j \mid \hat{\boldsymbol{z}}_S\right) \neq 0$. By continuous differentiability of p^{E_k}, we know that there exists $\delta_k^{(i,j,S)} > 0$ such that for all $\boldsymbol{M} \in \overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ such that $\|\boldsymbol{M} - \boldsymbol{I}\|_F \leqslant \delta_k^{(i,j,S)}$, the density of the variable $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}$ satisfies $q^{E_k}\left(\hat{\boldsymbol{v}}_i, \hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_S\right) \neq q^{E_k}\left(\hat{\boldsymbol{v}}_i \mid \hat{\boldsymbol{v}}_S\right) q^k\left(\hat{\boldsymbol{v}}_j \mid \hat{\boldsymbol{v}}_S\right)$ for $\hat{\boldsymbol{v}} = \boldsymbol{M}\hat{\boldsymbol{z}}$, which implies that $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are dependent given $\boldsymbol{v}_S$. Now choose $\delta = \min_{k,i,j,S} \delta_k^{(i,j,S)} > 0$, then for all $\boldsymbol{M} \in \overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ such that $\|\boldsymbol{M} - \boldsymbol{I}\|_F \leqslant \delta$, the resulting distributions $q^{E_k}, k \in [K]$ satisfy assumption Assumption 6.

Claim 2. There exists a neighbourhood O of $\boldsymbol{I}$ in $\overline{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ (in the sense of Remark 1) such that for almost all $\boldsymbol{M} \in O \cap \mathcal{M}_{\mathrm{sur}}^0(\mathcal{G})$, $p^{E_k}, k \in [K]$ satisfies Assumption 6 $\Rightarrow p^{E_k}, k \in [K]$ satisfies Assumption 6.

The proof is similar to the previous statement. Since Assumption 6 causal minimality is satisfied for $\boldsymbol{z}$, for $\forall k \in [K], i \in [d]$, let $\mathcal{G}_{ij}$ be the resulting graph obtained by removing the edge $j \to i$ from $\mathcal{G}$, then there must exists some $\alpha_{ijk} \in [d]$ such that $\boldsymbol{z}_{\alpha_{ijk}} \not\perp\!\!\!\perp \boldsymbol{z}_{\mathrm{nd}_{\mathcal{G}_{ij}}(\alpha_{ijk})} \mid \boldsymbol{z}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}$. Hence, there exists $\hat{\boldsymbol{z}}^{ijk} \in \mathbb{R}^d$ such that

$$p^{E_k}\left(\hat{\boldsymbol{z}}^{ijk}_{\alpha_{ijk}} \mid \hat{\boldsymbol{z}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right) p^{E_k}\left(\hat{\boldsymbol{z}}^{ijk}_{\mathrm{nd}_{\mathcal{G}_{ij}}(\alpha_{ijk})} \mid \hat{\boldsymbol{z}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right) \neq p^{E_k}\left(\hat{\boldsymbol{z}}^{ijk}_{\overline{\mathrm{nd}}_{\mathcal{G}_{ij}}(\alpha_{ijk})} \mid \hat{\boldsymbol{z}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right).$$

By continuous differentiability of p^{E_k}, there exists $\delta_k^{(i,j)} > 0$ such that for all $\boldsymbol{M} \in \bar{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ such that $\|\boldsymbol{M} - \boldsymbol{I}\|_F \leqslant \delta_k^{(i,j)}$, the density $q_{ij}^{E_k}$ of the variable $\hat{\boldsymbol{v}}^{ijk} = \boldsymbol{M}\hat{\boldsymbol{z}}^{ijk}$ satisfies

$$q^{E_k}\left(\hat{\boldsymbol{v}}^{ijk}_{\alpha_{ijk}} \mid \hat{\boldsymbol{v}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right) q^{E_k}\left(\hat{\boldsymbol{v}}^{ijk}_{\mathrm{nd}_{\mathcal{G}_{ij}}(\alpha_{ijk})} \mid \hat{\boldsymbol{v}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right) \neq q^{E_k}\left(\hat{\boldsymbol{v}}^{ijk}_{\overline{\mathrm{nd}}_{\mathcal{G}_{ij}}(\alpha_{ijk})} \mid \hat{\boldsymbol{v}}^{ijk}_{\mathrm{pa}_{\mathcal{G}_{ij}}(\alpha_{ijk})}\right).$$

for $\hat{\boldsymbol{v}}^{ijk} = \boldsymbol{M}\hat{\boldsymbol{z}}^{ijk}$. This implies that removing the edge $j \to i$ in $\mathcal{G}$ would break the causal Markov condition for q^{E_k}. Now let $\delta = \min_{k,i,j} \delta_k^{(i,j)} > 0$, then for all $\boldsymbol{M} \in \bar{\mathcal{M}}_{\mathrm{sur}}(\mathcal{G})$ such that $\|\boldsymbol{M} - \boldsymbol{I}\|_F \leqslant \delta$, the resulting distributions $q^{E_k}, k \in [K]$ satisfy assumption Assumption 1.

Combining the above two statements and what we have proven before, it is straightfoward to see that one can choose some $\boldsymbol{M} \in \mathcal{M}_{\mathrm{sur}}(\mathcal{G})$ in a small neighbourhood of $\boldsymbol{I}$ that satisfies all the requirements, completing the proof. □

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction provide the readers a sense of our main results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We compare with existing works in the introduction and the related work sections.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Rigorous proofs are provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We introduce our experimental setup in details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be released after review.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: we run experiments on 100 random causal graphs and report the overall accuracy.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments do not require huge computational resources and can be run on a local computer.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conforms the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.