Causal Representation Learning from General Environments under Nonparametric Mixing

Ignavier Ng[1] **Shaoan Xie**[1] **Xinshuai Dong**[1] **Peter Spirtes**[1] **Kun Zhang**[1,2]
[1]Carnegie Mellon University [2]Mohamed bin Zayed University of Artificial Intelligence

Abstract

Causal representation learning aims to recover the latent causal variables and their causal relations, typically represented by directed acyclic graphs (DAGs), from low-level observations such as image pixels. A prevailing line of research exploits multiple environments, which assume how data distributions change, including single-node interventions, coupled interventions, or hard interventions, or parametric constraints on the mixing function or the latent causal model, such as linearity. Despite the novelty and elegance of the results, they are often violated in real problems. Accordingly, we formalize a set of desiderata for causal representation learning that applies to a broader class of environments, referred to as general environments. Interestingly, we show that one can fully recover the latent DAG and identify the latent variables up to minor indeterminacies under a nonparametric mixing function and nonlinear latent causal models, such as additive (Gaussian) noise models or heteroscedastic noise models, by properly leveraging sufficient change conditions on the causal mechanisms up to third-order derivatives. These represent, to our knowledge, the first results to fully recover the latent DAG from general environments under nonparametric mixing. Notably, our results are stronger than many existing works, but require less restrictive assumptions about changing environments.

1 Introduction

Causal representation learning (CRL) aims to recover the latent causal variables and their causal structure from observations of variables that might be noncausal (Schölkopf et al., 2021). It is of great importance in many scientific fields, as in real-life complex systems, available measurements are often low-level, indirect, and high-dimensional, e.g., image pixels, linguistic tokens, and and gene expressions.

Despite growing interest, CRL remains a highly challenging task. Even in the case of independent latent variables—commonly referred to as nonlinear independent component analysis (ICA)—the problem is difficult (Hyvärinen et al., 2023). Nonlinear ICA, despite being a strictly easier task due to the lack of relationships among latent variables, is notoriously unidentifiable without additional assumptions, as different latent representations can explain the observed data equally well, yet may not align with the underlying data generating process (Hyvärinen and Pajunen, 1999). Furthermore, CRL inherits the complexities of causal discovery (Spirtes et al., 2001; Glymour et al., 2019), which is already challenging even when all causal variables are fully observed.

For the task of CRL, a line of work focuses on purely observational data with parametric and graphical assumptions. For instance, various graphical conditions have been proposed with linearity (Silva et al., 2006; Cai et al., 2019; Xie et al., 2020, 2022; Adams et al., 2021; Huang et al., 2022; Dong et al., 2023) or discrete assumptions (Kivva et al., 2021). Another prevailing line of research, which is the focus of this work, addresses the problem by making use of data from multiple distributions/environments, where the change of distribution is often assumed to arise from hard interventions (von Kügelgen et al., 2023; Jiang and Aragam, 2023; Bing et al., 2024), single-node interventions (Ahuja et al., 2023; Squires et al., 2023; Jiang and Aragam, 2023; Varici et al., 2023; Zhang et al., 2023), coupled interventions (von Kügelgen et al., 2023; Jin and Syrgkanis, 2023), or counterfactual views (Kügelgen et al., 2021; Brehmer et al., 2022). Other related works are further discussed in Appendix A.

Despite the novelty and elegance of the results, many of these assumptions regarding how distribution

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Table 1: Comparison of of several existing identifiability results of multi-environment CRL based on soft interventions. We only outline the key assumptions and results, while omitting some additional assumptions required by several studies. For a more comprehensive comparison, including works that rely on hard interventions, refer to Table 2 in the supplementary materials.

Work	Latent SEM	Mixing Function	Desiderata Satisfied?	Identifiability of Latent Variables	Identifiability of Latent DAG
Squires et al. (2023)	Linear	Linear	1,2	Up to ancestors	Transitive closure
Zhang et al. (2023)	Nonlinear	Polynomial	1,2	Up to ancestors	Transitive closure
Varıcı et al. (2024a)	General	Linear	1,2	Up to ancestors	Transitive closure
	Nonlinear	Linear	1,2	Up to surrounding	Full
Ahuja et al. (2023)	Bounded RV	Polynomial	1,2	Full	Full
Jin and Syrgkanis (2023)	General	**General**	1	Up to surrounding	Full
	Linear	Linear	**1,2,3**	Up to surrounding	Full
Varıcı et al. (2024b)	General	Linear	**1,2,3**	Up to ancestors	Transitive closure
Zhang et al. (2024)	General	**General**	**1,2,3**	Up to intimate neighbors	Moral graph
Ours	ANM	**General**	**1,2,3**	Up to intimate parents	Full
	HNM	**General**	**1,2,3**	Up to intimate parents	Full

changes are often violated in real problems. For instance, in genomics, it is commonplace to encounter soft interventions instead of hard interventions, e.g., in RNA interference (Dominguez et al., 2015). Accordingly, in this work, we formalize a set of desiderata to establish more realistic assumptions about changing environments for CRL, and we refer to the setting as *CRL from general environments*, detailed in Section 3. Under such a general and realistic CRL setting, we propose, to our knowledge, the first identifiability result to fully recover the latent DAG and identify the latent causal variables up to minor indeterminacies, while allowing for nonlinear latent causal models and nonparametric mixing. In contrast, the existing results for general environments either require linearity on the mixing function (Jin and Syrgkanis, 2023; Varıcı et al., 2024b), or can only identify the latent DAG up to its moral graph (Zhang et al., 2024). A comparison is provided in Table 1.

Concretely, we show that, with general environments, nonparametric mixing, and nonlinear latent causal models—such as additive (Gaussian) noise models (Theorem 1) or heteroscedastic noise models (Theorem 2)—it is possible to fully recover the latent DAG and identify the underlying latent variables up to minor indeterminacies. Specifically, each latent variable can be identified up to a mixture of itself and its *intimate parents* in the true latent DAG. While Zhang et al. (2024) exploit sufficient change conditions on the causal mechanisms up to second-order derivatives to identify the moral graph, we introduce a fundamentally different approach that properly leverages third-order derivatives to extract causal ordering information of the latent variables, enabling us to fully recover the latent DAG. This leads to a novel set of proof techniques, which we hope will inspire future research in this area. Finally, we validate our identifiability theory through simulation studies.

2 Problem Setting

Setup. We describe the problem setting of CRL. We assume that the observed random variables $X = (X_1, \ldots, X_d)$ and latent variables $Z = (Z_1, \ldots, Z_n)$ follow the data generating process below:

$$\begin{aligned} &\text{(Mixing)} && X = g(Z), \\ &\text{(Latent SEM)} && Z_i = f_i^{(u)}(\mathrm{PA}(Z_i; \mathcal{G}_Z), \epsilon_i^{(u)}),\ i \in [n]. \end{aligned} \tag{1}$$

Here, the observed random variables X are generated from the latent variables Z via an unknown, nonparametric mixing function g that is assumed to be a $\mathcal{C}^2$-diffeomorphism onto its image $\mathcal{X} \subseteq \mathbb{R}^d$. In each environment indexed by u, the latent variables Z follow a structural equation model (SEM), characterized the same but unknown DAG $\mathcal{G}_Z$ consisting of nodes $\{Z_i\}_{i=1}^n$. Here, $\mathrm{PA}(Z_i; \mathcal{G}_Z)$ denotes the parents of Z_i in DAG $\mathcal{G}_Z$.

Denote by $p_Z^{(u)}$ and $p_X^{(u)}$ the probability density functions of Z and X, respectively. To simplify notation and when the context is clear, we omit subscripts in the density functions and use X and Z to represent both the random variables and their specific values. We assume that $p_Z^{(u)}$ is third-order differentiable and has full support on $\mathbb{R}^n$.

A summary of the data generating process is illustrated in Fig. 1. In this work, we aim to estimate the latent variables Z and the latent DAG $\mathcal{G}_Z$ up to minor indeterminacies from samples of X across a number of

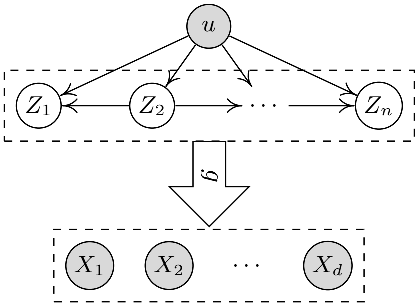

Figure 1: The observed random variables X are generated from the latent variables Z via an unknown, nonparametric mixing function g. The causal mechanism for each latent variable Z_i may vary across different environments specified by u. The gray shading of the nodes indicates that the variables are observable.

environments, i.e., $u \in [m]$. The distribution changes across environments may arise from heterogeneous data or nonstationary time series. Here, we focus on changes that can be represented by interventions—whether hard (perfect) or soft (imperfect) (Eberhardt, 2013; Yang et al., 2018).

Observational equivalence. For identification, we consider a model $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ that follows the data generating process specified in Eq. (1), and generates a data distribution that matches the distribution of the observed random variables X across the given environments, i.e.,

$$p_{\hat{X}}^{(u)}(x) = p_X^{(u)}(x), \quad \forall\, u \in [m],\, x \in \mathcal{X}^{(u)}, \tag{2}$$

where $\hat{Z}$ denote the estimated latent variables and $\hat{X} = \hat{g}(\hat{Z})$. In this case, the models $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ and $(g, p_Z, \mathcal{G}_Z)$ are said to be *observationally equivalent.* In practice, observational equivalence can be attained by maximum likelihood estimation in the large sample limit, where various approaches including normalizing flows (Rezende and Mohamed, 2015) can be applied.

Notations. We write $[n] \coloneqq \{1, \ldots, n\}$, and denote by $\oplus$ the vector concatenation symbol. For a matrix A, we denote by $A_{i,:}$ and $A_{:,j}$ its i-th row and j-th column, respectively. For a vector $Z = (Z_1, \ldots, Z_n)$, we write $Z_{[k]} \coloneqq (Z_1, \ldots, Z_k)$, and similarly for a matrix. With a slight abuse of notation, we occasionally treat a vector as a set. For example, we may write $Z_j \in Z$ if $Z_j \in \{Z_i\}_{i=1}^n$. Similar to the cardinality of a set, we denote by $|Z|$ the number of entries in vector Z. Moreover, we denote by $\mathcal{S}(\mathcal{G}_Z)$ the set of sink nodes in DAG $\mathcal{G}_Z$. For two DAGs $\mathcal{G}_Z$ and $\mathcal{G}_{\hat{Z}_\pi}$ where π is a permutation, we say that they are identical when $Z_i \to Z_j$ in $\mathcal{G}_Z$ if and only if $\hat{Z}_{\pi(i)} \to \hat{Z}_{\pi(j)}$ in $\mathcal{G}_{\hat{Z}}$. Following Zhang et al. (2024), we define the intimate neighbors of Z_i as

$$\Psi(Z_i; \mathcal{M}_Z) \coloneqq \{Z_j \mid Z_j, j \neq i, \text{ is adjacent to } Z_i \text{ and all other neighbors of } Z_i \text{ in } \mathcal{M}_Z\}.$$

3 Desiderata for CRL from General Environments

Nonlinear ICA may be viewed as a special case of CRL in which the true latent DAG is an empty graph. However, even in this case, the latent variables has shown to be unidentifiable without any assumptions (Hyvärinen and Pajunen, 1999). This challenge extends to CRL, a problem that is inherently more complex. As discussed in Section 1 and Appendix A, a prevailing line of research relies on distributional changes in the latent variables across different environments, including hard interventions, single-node interventions, or coupled interventions, some of which may be overly restrictive in practice.

In this work, we formalize a set of desiderata for CRL which applies to a broader class of environments that may be more realistic in real-world scenarios, referred to as general environments:

- **Desideratum 1:** The setting does not require hard interventions. The interventions can be either soft or hard.
- **Desideratum 2:** The setting does not assume any prior knowledge of intervention targets or their coupling pattern across environments.
- **Desideratum 3:** The interventions can be either single-node or multi-node across environments.

Note that the term "general environments" has been used by Jin and Syrgkanis (2023); here, we aim to provide a precise formulation of the desiderata, and use that term to refer to environments that satisfy the above desiderata. For Desideratum 1, hard interventions eliminate the dependency between the targeted variables and their parents, while soft interventions modify the causal mechanism of targeted variables without fully eliminating the influences of their parents, which thus are often considered more realistic (Yang et al., 2018; Jaber et al., 2020). For instance, in genomics, it is commonplace to encounter soft interventions, e.g., in CRISPR-mediated gene activation or RNA interference (Dominguez et al., 2015). For Desideratum 3, in real-world environments, the causal mechanisms of multiple latent variables may simultaneously change, resulting in multi-node interventions that extend beyond the scope of single-node interventions.

For Desideratum 2, since the latent variables are unknown by definition, it is in practice often infeasible to know the intervention targets and their coupling pattern (i.e., which environments share the same intervention targets), such as biology (Squires et al., 2020; Tejada-Lapuerta et al., 2023) and robotics (Lee et al., 2023). Furthermore, we show in Proposition 1 that knowing the coupling pattern of the intervention targets is equivalent to knowing the intervention targets (up to variable permutation) for single-node interventions, with a proof given in Appendix B.

Proposition 1 (Coupled single-node interventions)**.** *Consider a set of single-node interventions on the latent variables Z. The knowledge of which interventions share the same targets is equivalent to the knowledge of intervention targets up to variable permutation.*

In Tables 1 and 2 in the supplementary materials, we specify which desiderata above each existing result satisfies. Surprisingly, many existing results fail to satisfy at least one of the desiderata, indicating that current approaches may impose limitations or restrictive assumptions that hinder their applicability in more general, realistic environments.

4 CRL with Latent Additive Noise Models

In this section, we present our identifiability theory for general environments, which leverages specific properties of additive noise models (ANMs). Before discussing it, we first review the key idea of Zhang et al. (2024). In particular, the authors utilize the following property (Lin, 1997) regarding the conditional independence structure of the latent variables:

$$Z_i \perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}} \iff \frac{\partial^2 \log p^{(u)}(Z)}{\partial Z_i \partial Z_j} = 0. \qquad (3)$$

Denote by $\mathcal{M}_Z$ the Markov network over latent variables Z, where $\mathcal{M}_Z$ consists of nodes $\{Z_i\}_{i=1}^n$ and undirected edges $\{Z_i, Z_j\} \in \mathcal{E}(\mathcal{M}_Z)$ if and only if $Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}}$. Eq. (3) indicates that Z_i and Z_j are not adjacent in the Markov network if and only if the above equation holds. Zhang et al. (2024) then utilize sufficient changes of the latent distribution, involving second-order derivatives above, to recover the Markov network (or moral graph under the faithfulness assumption (Spirtes et al., 2001)) over Z in a nonparametric setting. However, since this approach essentially exploits a certain type of conditional independence information (that is invariant across all environments) via second-order derivatives that are symmetric, it generally cannot infer all causal directions, and can at most recover a Markov equivalence class over the latent variables, similar to constraint-based causal discovery approaches such as PC (Spirtes and Glymour, 1991).

In this work, we consider two classes of latent causal models and show that, interestingly, the third-order derivatives can reveal information about causal directions. This allows us to fully recover the latent DAG, by properly leveraging sufficient change conditions on the distributions up to third-order derivatives. Specifically, we focus on latent ANMs (Hoyer et al., 2009; Rolland et al., 2022) of the form:

$$Z_i = f_i^{(u)}(\mathrm{PA}(Z_i; \mathcal{G}_Z)) + \epsilon_i^{(u)}, \quad i \in [n], \qquad (4)$$

where the noise term ϵ_i follows Gaussian distribution and function $f_i^{(u)}$ is third-order differentiable. In the context of causal discovery from observed variables, it has been shown that the above causal model is identifiable under certain conditions (Hoyer et al., 2009; Rolland et al., 2022). We further discuss the use of heteroscedastic noise models in Section 5.

In Section 4.1, we discuss the property of sink nodes. We then provide our identifiability result in Section 4.2 and the estimation method in Section 4.3.

4.1 Third-Order Derivatives of Sink Nodes

We show how sink nodes provide information related to the third-order derivatives of the distribution of latent variables. Specifically, each sink node has the following property. The proof is given in Appendix C.1.

Lemma 1. *Consider the data generating process in Eq. (4) and let Z_i be a sink node in DAG $\mathcal{G}_Z$. Then,*

$$\frac{\partial \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0 \quad \textit{for} \quad j \in [d].$$

As discussed in Section 4.2, this property provides information that can be leveraged to identify the sink node from the latent Markov network. Once the sink node is identified, the Markov network property in Eq. (3) can be used to infer its parents. Specifically, the parents of a sink node Z_i are precisely the neighboring nodes of Z_i in the Markov network over Z.

It is worth noting that the above lemma closely resembles the work of Rolland et al. (2022), who developed a method for estimating causal structures among observed variables using score matching. However, they focus on causal discovery in settings where all variables are observed, and utilize the information that the *variance* of the second-order derivatives is zero. This limits its direct applicability to CRL. In contrast, as we will demonstrate in the next section, leveraging third-order derivatives as in Lemma 1 allows us to establish sufficient change conditions on the latent distribution, enabling identification of the latent DAG.

Moreover, in Lemma 1, we leverage the fact that sink node is a sufficient condition for the corresponding third-order partial derivative to be equal to zero. To establish it as a necessary condition, one could adopt similar assumptions to those in Rolland et al. (2022), such as requiring $f_i^{(u)}$ to be nonlinear in every component. It is worth noting that our results do not rely on this necessity, as long as Assumptions 1 and 2 can be satisfied.

4.2 Identifiability Theory

We introduce our identifiability result for general environments and nonparametric mixing. The key idea is to properly leverage sufficient change conditions on the causal mechanisms up to third-order derivatives, building upon Lemma 1.

We first define the following notion of *intimate parents* that is crucial for the identifiability result of the estimated latent variables:

$$\begin{aligned}\mathrm{IPA}(Z_i;\mathcal{G}_Z) \coloneqq \{&Z_j \in \mathrm{PA}(Z_i;\mathcal{G}_Z) \mid Z_j \text{ is a parent of}\\ &\text{every child of } Z_i\text{, and is either adjacent to,}\\ &\text{or a spouse of, every spouse of } Z_i \text{ in } \mathcal{G}_Z\}.\end{aligned}$$

Specifically, Zhang et al. (2024) showed that the underlying Markov network (or moral graph under faithfulness assumption) $\mathcal{M}_Z$ can be recovered, and that the latent variables can be identified up to permutation and mixture with intimate neighbors, i.e., $Z_i \cup \Psi(Z_i;\mathcal{M}_Z)$. In contrast, with the ANM described in Eq. (4), we show that, one can fully recover the latent DAG $\mathcal{G}_Z$ and identify the latent variables up to permutation and mixture with intimate parents, i.e., $Z_i \cup \mathrm{IPA}(Z_i;\mathcal{G}_Z)$. Importantly, we have $\mathrm{IPA}(Z_i;\mathcal{G}_Z) \subseteq \Psi(Z_i;\mathcal{M}_Z)$, indicating that our identifiability result significantly refines that of Zhang et al. (2024). Whereas their indeterminacy of latent variables involves intimate children and spouses, our result restricts the ambiguity to intimate parents only.

We now present the assumptions and the identifiability result, with a detailed explanation of both provided later in this section.

Assumption 1 (Sufficient changes for Markov network)**.** *Let Z' denote any subset of latent variables Z that includes all of their ancestors, and $n' \coloneqq |Z'|$. For each value of Z', there exist $2n' + |\mathcal{M}_{Z'}| + 1$ values of u, i.e., u_j with $j = 0, \ldots, 2n' + |\mathcal{M}_{Z'}|$, such that the vectors $\tau(Z',u_j) - \tau(Z',u_0)$ with $j = 1, \ldots, 2n' + |\mathcal{M}_{Z'}|$ are linearly independent, where vector $\tau(Z',u)$ is defined as:*

$$\begin{aligned}\tau(Z',u) = &\left(\frac{\partial \log p^{(u)}(Z')}{\partial Z_i}\right)_{Z_i\in Z'}\\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i^2}\right)_{Z_i\in Z'}\\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i \partial Z_j}\right)_{i<j\,:\,\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_{Z'})}.\end{aligned}$$

Assumption 2 (Sufficient changes for sink nodes)**.** *Let Z' denote any subset of latent variables Z that includes all of their ancestors. For each value of Z', there exist $L+1$ values of u, i.e., u_j with $j = 0, \ldots, L$, such that the vectors $w(Z',u_j) - w(Z',u_0)$ with $j = 1, \ldots, L$ are linearly independent, where $L \coloneqq |w(Z',u)|$ and vector $w(Z',u)$ is defined as:*

$$\begin{aligned}w(Z',u) = &\left(\frac{\partial \log p^{(u)}(Z')}{\partial Z_i}\right)_{Z_i\in Z'}\\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i^2}\right)_{Z_i\in Z'}\\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i \partial Z_j}\right)_{i<j\,:\,\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_{Z'})}\\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i^3}\right)_{Z_i\in Z'\setminus\mathcal{S}(\mathcal{G}_{Z'})}\\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i^2 \partial Z_j}\right)_{\substack{i<j\,:\,\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_{Z'}),\\ Z_i\notin\mathcal{S}(\mathcal{G}_{Z'})}}\\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i \partial Z_j \partial Z_k}\right)_{\substack{i<j<k\,:\,\{Z_i,Z_j\},\\ \{Z_j,Z_k\},\{Z_i,Z_k\}\in\mathcal{E}(\mathcal{M}_{Z'})}}.\end{aligned}$$

Theorem 1 (Identifiability with latent ANMs)**.** *Consider the data generating process in Eqs. (1) and (4). Suppose that Assumptions 1 and 2, as well as the faithfulness assumption, hold. Let $\mathcal{G}_{\hat{Z}}$ and $\hat{Z}$ be the output of Algorithm 1. Then, there exists a permutation π of the estimated latent variables $\hat{Z}$, denoted as $\hat{Z}_\pi$, such that:*

1. *(Identifiability of $\mathcal{G}_Z$) $\mathcal{G}_{\hat{Z}_\pi}$ and $\mathcal{G}_Z$ are identical.*
2. *(Identifiability of Z) $\hat{Z}_{\pi(i)}$ is solely a function of a subset of $\{Z_i\} \cup IPA(Z_i;\mathcal{G}_Z)$.*

The proof of the theorem is provided in Appendix D, which makes use of Lemma 1. Here, the word 'solely' implies that $\hat{Z}_{\pi(i)}$ is not a function of other latent variables Z_j not specified above. We now outline the key intuitions behind the assumptions and the identifiability result. Intuitively speaking, Assumptions 1 and 2 impose certain type of sufficient change conditions on the distribution of latent variables. Note that different forms of sufficient change conditions have been

Algorithm 1 Iterative identification of Markov network and sink nodes

```
Input: $p^{(u)}(X), u = 1, \ldots, m$
Output: $\mathcal{G}_{\hat{Z}}$ and $\hat{Z}$
1: for $t = 1, \ldots, n-1$ do
2:     Learn $(\hat{g}^t, p^t_{\hat{Z}}, \mathcal{G}^t_{\hat{Z}})$ to achieve Eq. (2) with minimal number of edges for Markov network $\mathcal{M}^t_{\hat{Z}_{[n+1-t]}}$, while (i) constraining the local structure (including incoming and outgoing edges) of $\{\hat{Z}_i\}_{i=n+2-t}^{n}$ in $\mathcal{G}^t_{\hat{Z}}$ to be identical to those in $\mathcal{G}^{t-1}_{\hat{Z}}$, and (ii) ensuring $\mathcal{M}^t_{\hat{Z}_{[k]}} \subseteq \mathcal{M}^{t-1}_{\hat{Z}_{[k]}}$ for $k = n+2-t, \ldots, n$.
3:     Let $\hat{Z}_i$ a sink node in DAG $\mathcal{G}^t_{\hat{Z}_{[n+1-t]}}$. Reorder $\hat{Z}$ in $(\hat{g}^t, p^t_{\hat{Z}}, \mathcal{G}^t_{\hat{Z}})$ by swapping $\hat{Z}_i$ and $\hat{Z}_{n+1-t}$.
4: return $\mathcal{G}^{n-1}_{\hat{Z}}$ and $(\hat{g}^{n-1})^{-1}(X)$
```

adopted in nonlinear ICA (Hyvärinen et al., 2023) and CRL (Zhang et al., 2024).

First, under Assumption 1, observational equivalence, together with a sparsity constraint on the recovered Markov network, allows us to recover the true Markov network (or moral graph under faithfulness assumption) up to isomorphism and to identify the latent variables up to permutation and mixture with their intimate neighbors, similar to the result of Zhang et al. (2024). Next, going beyond Markov network, Assumption 2 enables us to further identify the sink nodes of the DAG, whose parents are precisely their corresponding neighbors in the recovered Markov network. At this stage, having identified one of the sink nodes and its parents, we can fix the local structure of the sink node, and proceed to learn the edges among the remaining nodes, while ensuring that the Markov network (or moral graph) remains consistent with the previous iterations.[1] This iterative procedure allows us to progressively identify sink nodes and their local structure at each step, and move to the next stage. Ultimately, this process leads to the identification of the entire DAG, as well as the latent variables up to mixture with their intimate parents.

The resulting method is described in Algorithm 1. Due to the permutation indeterminacy of the latent variables, we design Algorithm 1 to output $\hat{Z}$ in a causal order with respect to the latent DAG $\mathcal{G}_{\hat{Z}}$. It is worth noting that the overall idea is analogous to ordering-based causal discovery methods (Raskutti and Uhler, 2018; Ghoshal and Honorio, 2018; Rolland et al., 2022), which aim to identify the causal order assuming that all causal variables are observed. However, since the relevant causal variables are unobserved in our setting, our result requires careful reasoning over the latent variables and latent DAG, leading to entirely different assumptions and algorithmic procedure.

Example 1 (Identifiability result)**.** *Consider the true latent DAG $\mathcal{G}_Z : Z_1 \rightarrow Z_2 \leftarrow Z_3$. Applying Theorem 1, there exist a permutation π of the estimated latent variables $\hat{Z}$, such that $\mathcal{G}_{\hat{Z}_\pi}$ and $\mathcal{G}_Z$ are identical. Also, we have: (a) $\hat{Z}_{\pi(1)}$ is solely a function of Z_1, (b) $\hat{Z}_{\pi(2)}$ is solely a function of $\{Z_1, Z_2, Z_3\}$, and (c) $\hat{Z}_{\pi(3)}$ is solely a function of Z_3. Here, both Z_1 and Z_3 are identified up to component-wise transformation.*

In the example above, the identifiability theory by Zhang et al. (2024) can only recover the moral graph of $\mathcal{G}_Z$, i.e., an undirected triangle. Moreover, in their result, each estimated latent variable $\hat{Z}_{\pi(i)}$ can still be a function of all latent variables $\{Z_1, Z_2, Z_3\}$, indicating that no disentanglement is achieved. In contrast, our identifiability theory significantly advances upon their result by providing a stronger recovery of the latent causal structure and latent variables.

Additionally, Jin and Syrgkanis (2023) investigated the setting of single-node and coupled soft interventions, showing that the true latent DAG can be fully recovered, with the latent variables identified up to permutation and mixture with *surrounding nodes*. Our identifiability result is stronger than theirs, despite not requiring either single-node interventions or coupled interventions. Specifically, we identify each latent variables up to mixture with its intimate parents, whose indeterminacy is smaller than theirs involving surrounding nodes. The reason is that our result also leverages sufficient change conditions w.r.t the Markov network (e.g., see Assumption 1).

Number of environments. Assumptions 1 and 2 require multiple environments to satisfy the assumption of linear independence across partial derivatives of latent distributions. Denote by $|\Delta(\mathcal{M}_Z)|$ be the number of triangles in the Markov network (or moral graph under the faithfulness assumption) over latent variables Z. The number of environments we need is (at least) $3n + 3|\mathcal{E}(\mathcal{M}_Z)| + |\Delta(\mathcal{M}_Z)| - 2|\mathcal{S}(\mathcal{G}_Z)| + 1$. Note that Zhang et al. (2024) require (at least) $2n + |\mathcal{E}(\mathcal{M}_Z)| + 1$ environments for recovering the latent Markov network, while we leverage more environments to recover the latent DAG.

Discussion of assumptions. Intuitively speaking, Assumptions 1 and 2 requires that the causal mechanisms among the latent variables change sufficiently across different environments. Such distribution changes, together with the invariant mixing function,

[1] For clarity, in Lines 2 and 3 of Algorithm 1, we handle only one sink node per iteration, even though multiple sink nodes may exist.

provide useful information for us to recover the latent variables and their causal relations.

To illustrate, consider a model with three latent variables that follow the SEM: $Z_1 = \epsilon_1^{(u)}$, $Z_2 = f_2^{(u)}(Z_1) + \epsilon_3^{(u)}$, and $Z_3 = f_3^{(u)}(Z_2) + \epsilon_2^{(u)}$, where functions $f_2^{(u)}$ and $f_3^{(u)}$ are parameterized by multilayer perceptrons (MLPs) (which can approximate any function). Here, the latent DAG is $Z_1 \to Z_2 \to Z_3$. Suppose that the variances of the noise terms $\epsilon_i^{(u)}$ and the weights of the MLPs are generated randomly for different environment u. Then, the sufficient change assumptions will hold almost surely.

Furthermore, Assumptions 1 and 2 support both soft and hard interventions, as well as single-node or multi-node interventions. These assumptions also do not assume prior knowledge of intervention targets or their coupling patterns. Therefore, the desiderata described in Section 3 can be satisfied.

To discuss some implications, our assumptions imply that the functions $f_i^{(u)}, i \in [n]$ cannot be linear in all environments. Otherwise, all third-order derivatives involved in Assumption 2 will vanish across environments, rendering the assumption unsatisfiable. Loosely speaking, this suggests that nonlinear functions may sometimes offer greater potential for satisfying such assumptions.

With nonparametric mixing. Existing results addressing nonparametric mixing often assume that the coupling pattern of single-node interventions is known, i.e., which interventions target the same latent variables (von Kügelgen et al., 2023; Jin and Syrgkanis, 2023). This assumption arises in part because their proofs rely primarily on first-order derivatives of the distribution of latent variables, necessitating knowledge of which causal mechanisms change together to establish sufficient changes on the distributions of latent variables. Zhang et al. (2024) improved on this by leveraging second-order derivatives to recover the Markov network without requiring knowledge of the coupling pattern. In this work, we utilize third-order derivatives to achieve stronger identifiability results, allowing us to fully recover the latent DAG without assuming prior knowledge of the coupling pattern.

4.3 Estimation Method

Building on the identifiability result in Theorem 1, we now develop a practical method for Algorithm 1. Specifically, Line 2 of Algorithm 1 requires achieving observational equivalence in Eq. (2), which involves maximum likelihood estimation. Here, we use variational autoencoder (VAE) (Kingma and Welling, 2014), following previous works (Khemakhem et al., 2020; Zhang et al., 2024). Other estimation methods can also be adopted, such as normalizing flows (Rezende and Mohamed, 2015; Dinh et al., 2017).

Specifically, for Line 2 of Algorithm 1, we maximize the marginal likelihood for each environment as

$$\begin{aligned}
&\log p^{(u)}(X) \\
&= \log \int_{\hat{Z}} p^{(u)}(X|\hat{Z}) p^{(u)}(\hat{Z}) d\hat{Z} \\
&= \log \int_{\hat{Z}} \frac{q^{(u)}(\hat{Z}|X)}{q^{(u)}(\hat{Z}|X)} p^{(u)}(X|\hat{Z}) p^{(u)}(\hat{Z}) d\hat{Z} \\
&\geq \mathbb{E}_{q^{(u)}(\hat{Z}|X)}(\log p^{(u)}(X|\hat{Z})) - D_{KL}(q^{(u)}(\hat{Z}|X), p^{(u)}(\hat{Z})) \\
&= -\mathcal{L}_{\text{elbo}},
\end{aligned}$$

where second term denotes the Kullback–Leibler (KL) divergence between between the prior distribution $p^{(u)}(\hat{Z})$ and the approximate posterior distribution $q^{(u)}(\hat{Z}|X)$. Using standard proof technique (Khemakhem et al., 2020; Lachapelle et al., 2024), we can show that if the posterior $q^{(u)}(\hat{Z}|X)$ has enough capacity to express the true posterior, then Eq. (2) can be satisfied. Following existing works, we assume that the posterior $q^{(u)}(\hat{Z}|X)$ is a multivariate Gaussian distribution, where its mean and diagonal covariance are output by an encoder that is modeled as a MLP. The decoder, also modeled as an MLP, outputs the mean of $p^{(u)}(X|\hat{Z})$ that is assumed to be a multivariate Gaussian distribution, with a pre-specified variance, and thus the first term reduces to the reconstruction error of data from estimated latent variables $\hat{Z}$. Note that this decoder corresponds to the mixing function $\hat{g}$.

Now we need to construct a proper prior distribution $p^{(u)}(\hat{Z})$ to recover the dependent latent variables. Following Eq. (4), we model $\log p^{(u)}(\hat{Z})$ using the Gaussian log-likelihood:

$$\begin{aligned}
\log p^{(u)}(\hat{Z}) = &-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\hat{Z}_i - \hat{f}^{(u)}(\hat{A}_{i,:}\hat{Z}) - \hat{\mu}_i^{(u)}}{\hat{\sigma}_i^{(u)}}\right)^2 \\
&-\frac{1}{2}\sum_{i=1}^{n}\log(\hat{\sigma}_i^{(u)})^2,
\end{aligned}$$

where where $\hat{f}_i^{(u)}$ is constructed as a three-layer MLP and $\hat{A}$ is a learnable binary adjacency matrix that represents the estimated DAG $\mathcal{G}_{\hat{Z}}$. Specifically, $\hat{A}_{i,j} = 1$ indicates $\hat{Z}_j \to \hat{Z}_i$. We restrict $\hat{A}$ to be strictly lower triangular to avoid cyclic causal relations. Furthermore, we learn $\hat{\mu}_i^{(u)}$ and $\hat{\sigma}_i^{(u)}$ via $\hat{\mu}^{(u)} = h_i^{\mu}(u)$ and $\hat{\sigma}_i^{(u)} = h_i^{\sigma}(u)$, where h_i^{μ} and h_i^{σ} are MLPs. Furthermore, in Algorithm 1, we can find the sink nodes according to binary adjacency matrix $\hat{A}$, and constrain the local structure (which corresponds to constraint (i) in Line 2 of Algorithm 1) by fixing the corresponding parameters in the adjacency matrix $\hat{A}$ to certain binary values and further restricting them to be non-learnable during the training process.

We now discuss how to handle constraint (ii) in Line 2 of Algorithm 1. Denote by A^{t-1} the binary adjacency matrix estimated in the previous (i.e., $(t-1)$-th) iteration. Let $\mathcal{L}_{\text{moral}}$ be defined as

$$\sum_{k=n+2-t}^{n} \big\|(I+\hat{A}_{[k],[k]})^T(I+\hat{A}_{[k],[k]}) - (I+A^{t-1}_{[k],[k]})^T(I+A^{t-1}_{[k],[k]})\big\|_F^2.$$

Instead of a hard constraint $\mathcal{L}_{\text{moral}}=0$ which can be enforced by constrained optimization methods, we simply treat it as a regularization term. Moreover, to enforce sparsity, we impose a constraint on the moral graph via the following regularization term, i.e., $\mathcal{L}_{\text{sparsity}} = \|(I+\hat{A})^T(I+\hat{A})\|_1$. The full objective function becomes $\mathcal{L}_{\text{elbo}}+\lambda_1\mathcal{L}_{\text{sparsity}}+\lambda_2\mathcal{L}_{\text{moral}}$.

After Algorithm 1 ends, the outputs of the encoder correspond to the estimated latent variables $\hat{Z}$ from the observations X, while the binary matrix $\hat{A}$ represents the recovered DAG $\mathcal{G}_{\hat{Z}}$.

5 CRL with Latent Heteroscedastic Noise Models

After presenting the identifiability result for ANMs in Section 4, we now consider heteroscedastic noise models (HNMs) (Xu et al., 2022; Immer et al., 2023) that may be more general, where the noise terms are not assumed to have constant variances:

$$Z_i = f_i^{(u)}(\text{PA}(Z_i;\mathcal{G}_Z))+\sigma_i^{(u)}(\text{PA}(Z_i;\mathcal{G}_Z))\epsilon_i^{(u)}, \quad i\in[n]. \tag{5}$$

Here, the noise term ϵ_i follows Gaussian distribution, and functions $f_i^{(u)}$ and $\sigma_i^{(u)}$ are third-order differentiable. In the context of causal discovery from observed variables, it has been shown that the above causal model is identifiable under certain conditions (Xu et al., 2022; Immer et al., 2023). The estimation method for HNMs is nearly identical to that for ANMs in Section 4.3, which we omit here.

5.1 Third-Order Derivatives of Sink Nodes

Analogous to Lemma 1 for ANMs, we present a corresponding result for HNMs. Specifically, the third-order derivatives of the sink node exhibit a property that provides valuable information for inferring the latent DAG. The proof is provided in Appendix C.2.

Lemma 2. *Consider the data generating process in Eq. (5) and let Z_i be a sink node in DAG $\mathcal{G}_Z$. Then,*

$$\frac{\partial \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0 \quad \textit{for} \quad Z_j \notin PA(Z_i;\mathcal{G}_Z).$$

5.2 Identifiability Theory

We now present the identifiability theory for latent HNMs in general environments under nonparametric mixing. As with the setting of ANMs, the key idea is to properly leverage sufficient change conditions on the causal mechanisms of latent variables up to third-order derivatives, building on Lemma 2.

The assumptions and results are presented below, with a proof provided in Appendix E. The assumptions and results here closely mirror those for ANMs, and we refer readers to Section 4.2 for a detailed explanation.

Assumption 3 (Sufficient changes for sink nodes)**.** *Let Z' denote any subset of latent variables Z that includes all of their ancestors. For each value of Z', there exist $L+1$ values of u, i.e., u_j with $j=0,\ldots,L$, such that the vectors $w(Z',u_j)-w(Z',u_0)$ with $j=1,\ldots,L$ are linearly independent, where $L:=|w(Z',u)|$ and vector $w(Z',u)$ is defined as:*

$$\begin{aligned} w(Z',u) = &\left(\frac{\partial \log p^{(u)}(Z')}{\partial Z_i}\right)_{Z_i\in Z'} \\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i^2}\right)_{Z_i\in Z'} \\ &\oplus \left(\frac{\partial^2 \log p^{(u)}(Z')}{\partial Z_i \partial Z_j}\right)_{i<j\,:\,\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_{Z'})} \\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i^3}\right)_{Z_i\in Z'\setminus\mathcal{S}(\mathcal{G}_{Z'})} \\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i^2 \partial Z_j}\right)_{i<j\,:\,\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_{Z'})} \\ &\oplus \left(\frac{\partial^3 \log p^{(u)}(Z')}{\partial Z_i \partial Z_j \partial Z_k}\right)_{\substack{i<j<k\,:\,\{Z_i,Z_j\},\\ \{Z_j,Z_k\},\{Z_i,Z_k\}\in\mathcal{E}(\mathcal{M}_{Z'})}}. \end{aligned}$$

Theorem 2 (Identifiability with latent HNMs)**.** *Consider the data generating process in Eqs. (1) and (5). Suppose that Assumptions 1 and 3, as well as the faithfulness assumption, hold. Let $\mathcal{G}_{\hat{Z}}$ and $\hat{Z}$ be the output of Algorithm 1. Then, there exists a permutation π of the estimated latent variables $\hat{Z}$, denoted as $\hat{Z}_\pi$, such that:*

1. *(Identifiability of $\mathcal{G}_Z$) $\mathcal{G}_{\hat{Z}_\pi}$ and $\mathcal{G}_Z$ are identical.*
2. *(Identifiability of Z) $\hat{Z}_{\pi(i)}$ is solely a function of a subset of $\{Z_i\}\cup IPA(Z_i;\mathcal{G}_Z)$.*

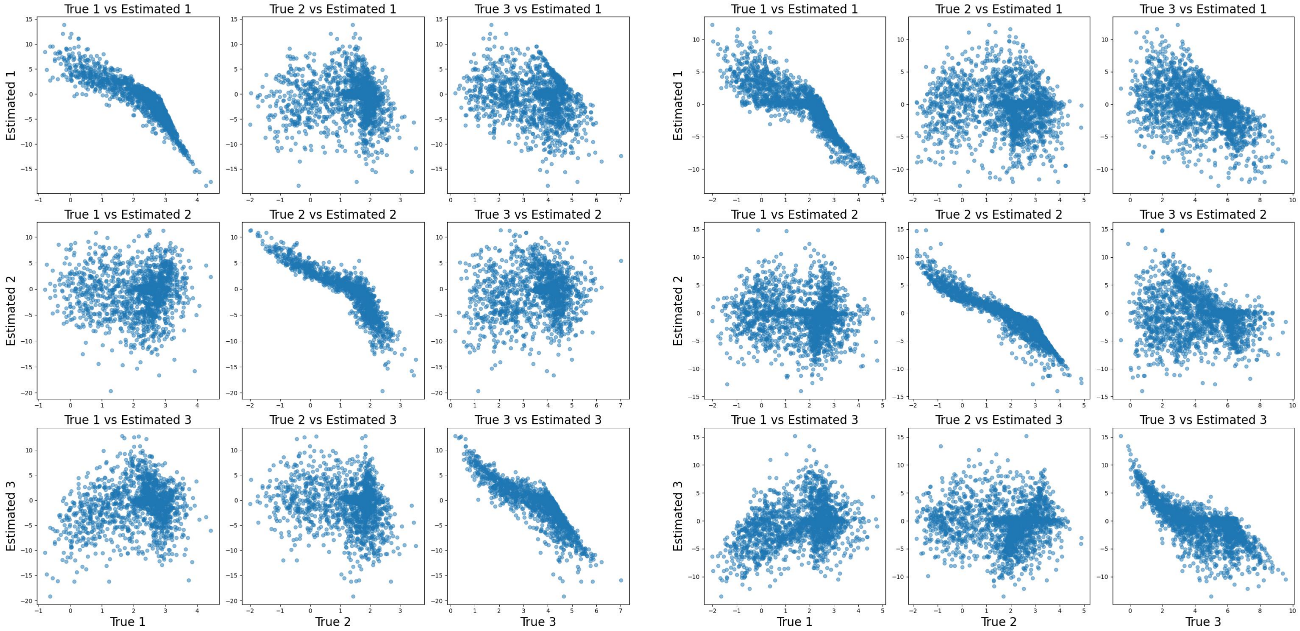

Figure 2: The recovered latent variables $\hat{Z}$ versus the true latent variables Z under two latent DAGs: (1) $Z_1 \to Z_2 \to Z_3$ and (2) $Z_1, Z_2 \to Z_3$.

6 Simulation Studies

We conduct simulation studies to verify our identifiability theory. We consider three latent variables that follow our predefined DAG and the data generating procedure in Eq. (4). The mean and variances of $\epsilon_i^{(u)}$ are randomly sampled across different environments, and $f_i^{(u)}$ is modeled as a two-layer MLP that consists of weights randomly sampled from $\text{Unif}[-2, 2]$ and LeakyReLU activation function. For the mixing function, we employ a two-layer MLP to transform the latent variables Z into observed variables X, ensuring injectivity by constraining the weight matrix to be orthogonal and using the LeakyReLU activation function. It is worth noting that the data generating procedure considered here satisfies the desiderata described in Section 3. During the estimation process, we employ three-layer MLPs as the encoder and decoder of VAE. We use the Adam optimizer (Kingma and Ba, 2014) to train the model with learning rate of 0.001.

The results are shown in Fig. 2, where each subfigure is a 3×3 panel. The panel on i-th row and j-th column displays the relationship between the estimated latent variable $\hat{Z}_i$ and the true latent variable Z_j. Our method successfully recovers the correct structure in both cases. Notably, there is a clear one-to-one correspondence between Z_1 and $\hat{Z}_1$, as well as between Z_2 and $\hat{Z}_2$, indicating that these variables are identified up to component-wise transformations. The correspondence between Z_3 and $\hat{Z}_3$ is less pronounced, which is expected since $\hat{Z}_3$ is recovered as a function of both Z_3 and its intimate parent Z_2. These observations validate our identifiability theory in Theorem 1. Further empirical studies on selecting number of latent variables and the hyperparameters are provided in Appendices F.1 and F.2, respectively.

7 Conclusion

We formalize a set of desiderata to establish more realistic assumptions about changing environments in CRL, which we refer to as general environments. Under this setting, we show that it is possible to fully recover the latent DAG and identify the latent variables up to minor indeterminacies, while allowing for nonparametric mixing and nonlinear latent causal models, such as ANMs and HNMs with Gaussian noise. Our results properly leverage sufficient change conditions on the causal mechanisms up to third-order derivatives, to extract causal ordering information and fully recover the latent DAG. We validate our identifiability theory through simulation studies. Future works include applying the proposed method to real-world data and further relaxing the assumption of Gaussian noise.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments. We would like to acknowledge the support from NSF Award No. 2229881, AI

Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program. IN acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarships – Doctoral program.

References

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

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, 2023.

A. Ben-Israel. The change-of-variables formula using matrix volume. *Siam Journal on Matrix Analysis and Applications*, 21, 01 1999.

S. Bing, U. Ninad, J. Wahl, and J. Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Conference on Causal Learning and Reasoning*, 2024.

J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, 2022.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. K. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Conference on Neural Information Processing Systems*, 2023.

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

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

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

A. Dominguez, W. Lim, and L. Qi. Beyond editing: Repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nature reviews. Molecular cell biology*, 17, 12 2015.

X. Dong, B. Huang, I. Ng, X. Song, Y. Zheng, S. Jin, R. Legaspi, P. Spirtes, and K. Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *The Twelfth International Conference on Learning Representations*, 2023.

F. Eberhardt. Direct causes and the trouble with soft interventions. *Erkenntnis*, 79, 08 2013.

N. Friedman and D. Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50: 95–125, 2003.

M. C. Gemici, D. Rezende, and S. Mohamed. Normalizing flows on Riemannian manifolds. *arXiv preprint arXiv:1611.02304*, 2016.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, 2018.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

H. Hälvä and A. Hyvärinen. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. *Advances in Neural Information Processing Systems*, 34, 2021.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, 2009.

B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, 35:5549–5561, 2022.

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

A. Hyvärinen, I. Khemakhem, and H. Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023. ISSN 2666-3899.

A. Immer, C. Schultheiss, J. E. Vogt, B. Schölkopf, P. Bühlmann, and A. Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, 2023.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, 2020.

Y. Jiang and B. Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

J. Jin and V. Syrgkanis. Learning causal representations from general environments: Identifiability and intrinsic ambiguity. *arXiv preprint arXiv:2311.12267*, 2023.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

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

A. Kori, P. Sanchez, K. Vilouras, B. Glocker, and S. A. Tsaftaris. A causal ordering prior for unsupervised representation learning. *arXiv preprint arXiv:2307.05704*, 2023.

J. V. Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

S. Lachapelle, P. R. López, Y. Sharma, K. Everett, R. L. Priol, A. Lacoste, and S. Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

W.-Y. Lam, B. Andrews, and J. Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

T. E. Lee, S. Vats, S. Girdhar, and O. Kroemer. SCALE: Causal learning and discovery of robot manipulation skills using simulation. In *7th Annual Conference on Robot Learning*, 2023.

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. Causal component analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

J. Lin. Factorizing multivariate function classes. *Advances in neural information processing systems*, 10, 1997.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and S. Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, 2022.

Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. van den Hengel, K. Zhang, and J. Q. Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2023.

Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. van den Hengel, K. Zhang, and J. Q. Shi. Identifiable latent polynomial causal models through the lens of change. In *International Conference on Learning Representations*, 2024.

F. Montagna, A. Mastakouri, E. Eulig, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, and F. Locatello. Assumption violations in causal discovery and the robustness of score matching. In *Advances in Neural Information Processing Systems*, 2023a.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, 2023b.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 752–771, 2023c.

H. Morioka and A. Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *International Conference on Machine Learning*, 2024.

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

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on International Conference on Machine Learning*, 2015.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, 2022.

P. Sanchez, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations*, 2023.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(8): 191–246, 2006. URL http://jmlr.org/papers/v7/silva06a.html.

L. Solus, Y. Wang, and C. Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). *arXiv preprint arXiv:2001.04872*, 2020.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT press, 2nd edition, 2001.

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

C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, 2023.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio, and F. J. Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

B. Varici, E. Acarturk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning: Linear and general transformations. *arXiv preprint arXiv:2402.00849*, 2024a.

B. Varıcı, E. Acartürk, K. Shanmugam, and A. Tajer. Linear causal representation learning from unknown multi-node interventions. *arXiv preprint arXiv:2406.05937*, 2024b.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

J. von Kügelgen, M. Besserve, L. Wendong, L. Gresele, A. Kekić, E. Bareinboim, D. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

Y. Wang and M. I. Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

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

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

D. Xu, D. Yao, S. Lachapelle, P. Taslakian, J. von Kügelgen, F. Locatello, and S. Magliacane. A sparsity principle for partially observable causal representation learning. In *International Conference on Machine Learning*, 2024.

S. Xu, A. Marx, O. Mian, and J. Vreeken. Causal inference with heteroscedastic noise models. In *Proceedings of the AAAI Workshop on Information Theoretic Causal Inference and Discovery*, 2022.

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

M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024.

W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2022.

J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 2023.

K. Zhang, S. Xie, I. Ng, and Y. Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning*, 2024.

Y. Zheng and K. Zhang. Generalizing nonlinear ICA beyond structural sparsity. *Advances in Neural Information Processing Systems*, 2023.

Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The assumptions are clearly stated in each theorem statement.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The primary focus of this work is on identifiability theory.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The assumptions are clearly stated in each theorem statement.
   (b) Complete proofs of all theoretical results. [Yes] The proofs are provided in the supplementary materials.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] The details are explained in Section 6.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendices

A Extended Discussion of Related Works

Nonlinear ICA. Despite the lack of dependence among latent variables, nonlinear ICA is a challenging problem because the latent variables are generally not identifiable without any further assumptions (Hyvärinen and Pajunen, 1999). Existing work in nonlinear ICA often relies on assumptions about changing distributions where different distributions are indicated by auxiliary variables such as time indices and domain indices (Hyvarinen and Morioka, 2016, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020; Lachapelle et al., 2024; Hälvä and Hyvärinen, 2020; Hälvä et al., 2021; Yao et al., 2022) or specific constraints on the mixing function, including function classes (Hyvärinen and Pajunen, 1999; Taleb and Jutten, 1999; Gresele et al., 2021; Buchholz et al., 2022) or sparsity constraint (Zheng et al., 2022; Zheng and Zhang, 2023).

Causal representation learning. CRL aims to recover the latent variables and their causal relations from data. Without any further assumptions, the identifiability of the hidden generating process is known to be impossible. One line of work focuses on purely observational data by adding parametric and graphical assumptions. In the linear case, Silva et al. (2006) recover the Markov equivalence class for the one-factor model, while Xie et al. (2020); Cai et al. (2019) estimate the latent variables and their relations by assuming non-Gaussian noises and at least twice measured variables as pure children as latent ones, which is later extended to latent hierarchical structure (Xie et al., 2022). Moreover, Adams et al. (2021) provides necessary and sufficient graphical conditions for the identification of linear non-Gaussian model with latent variables. As for the linear Gaussian case, Huang et al. (2022) leverage rank deficiency constraints of sub-covariance matrix over observed variables to identify latent hierarchical structure, which is later extended by Dong et al. (2023) to structures where all variables including latent and observed ones are allowed to be flexibly related. In the discrete setting, Kivva et al. (2021) also provide identification results for latent causal graphs up to Markov equivalence, by assuming a mixture model with specific graphical assumptions.

Due to the difficulty of using purely observational data for CRL, researchers have also been working on another line of thought where data from multiple environments are available and the changes of distributions are often assumed to arise from soft or hard interventions. Specifically, Squires et al. (2023) assume linear SEM and mixing function and leverage single-node interventions for the identifiability, Varici et al. (2023) consider nonlinear SEM and linear mixing function, Varici et al. (2023); Buchholz et al. (2023); Jiang and Aragam (2023) concern the setting of the nonparametric latent SEM and linear mixing function, which is further generalized by Ahuja et al. (2023) to nonparametric SEM and polynomial mixing functions, and Brehmer et al. (2022); von Kügelgen et al. (2023); Jiang and Aragam (2023) consider the general nonparametric settings for both the causal model and mixing function. From the perspective of assumptions about interventions, Squires et al. (2023) make use of single-node interventions, Brehmer et al. (2022) rely on counterfactual pre- and post-intervention views, von Kügelgen et al. (2023) utilize paired interventions per node, and Zhang et al. (2023) explore soft interventions. Other works include Yang et al. (2021); Shen et al. (2022); Liang et al. (2023) that require more supervision information, Kori et al. (2023) that require a causal ordering prior, Yao et al. (2024); Xu et al. (2024) that focus on multi-view data, Morioka and Hyvärinen (2024) that leverage certain graphical constraint, Ahuja et al. (2023); Wang and Jordan (2021) that utilize further constraint on the latent support, and Lippe et al. (2022) that provides results for interventions on known targets.

At the same time, many of the existing results rely on assumptions about interventions that may be overly restrictive in practice. As such, we formalize a set of desiderata about realistic assumptions about changing environments, which we refer to as CRL from general environments. Under this setting, we propose the first identifiability result that allows nonlinear SEM and nonparametric mixing, while the two strongest existing works either have to require linear mixing (Varıcı et al., 2024b), or can only identify the latent DAG up to moral graph (Zhang et al., 2024).

Ordering-based causal discovery. The idea that a causal DAG can be partially represented by a topological ordering has a long history (Verma and Pearl, 1990). Such a line of thought typically searches the ordering space

Table 2: Comparison of of several existing identifiability results of multi-environment CRL based on hard or soft interventions. We only outline the key assumptions and results, while omitting some additional assumptions required by several studies. 'Env.' stands for environments.

Work	Latent SEM	Mixing Function	Desiderata Satisfied?	General Env.?	Identifiability of Latent Variables	Identifiability of Latent DAG
Liu et al. (2023)	General	Linear	3	No	Full	Full
Squires et al. (2023)	Linear	Linear	2	No	Full	Full
	Linear	Linear	1,2	No	Up to ancestors	Transitive closure
Buchholz et al. (2023)	Linear	General	2	No	Full	Full
	Linear	General	1,2	No		Partial order
Zhang et al. (2023)	Nonlinear	Polynomial	1,2	No	Up to ancestors	Transitive closure
Liu et al. (2024)	Polynomial	General	3	No	Full	Full
Varıcı et al. (2024a)	General	Linear	2	No	Full	Full
	General	Linear	1,2	No	Up to ancestors	Transitive closure
	Nonlinear	Linear	1,2	No	Up to surrounding	Full
	General	**General**	2	No	Full	Full
Ahuja et al. (2023)	General	Polynomial	2	No	Full	Full
	Bounded RV	Polynomial	1,2	No	Full	Full
von Kügelgen et al. (2023)	General	**General**	None	No	Full	Full
Jin and Syrgkanis (2023)	General	**General**	1	No	Up to surrounding	Full
	Linear	Linear	1,2,3	**Yes**	Up to surrounding	Full
Bing et al. (2024)	General	Linear	2,3	No	Full	Full
Varıcı et al. (2024b)	ANM	Linear	2,3	No	Full	Full
	General	Linear	1,2,3	**Yes**	Up to ancestors	Transitive closure
Zhang et al. (2024)	General	**General**	1,2,3	**Yes**	Up to intimate neighbors	Moral graph
Ours	ANM	**General**	1,2,3	**Yes**	Up to intimate parents	Full
	HNM	**General**	1,2,3	**Yes**	Up to intimate parents	Full

instead of the DAG space, e.g., with greedy Markov Chain Monte Carlo (Friedman and Koller, 2003), greedy hill-climbing (Teyssier and Koller, 2012), restricted maximum likelihood estimators (Teyssier and Koller, 2012), and sparsest permutation (Raskutti and Uhler, 2018; Lam et al., 2022; Solus et al., 2021). Further, assumptions about functional causal models and noise variances can also be made (Ghoshal and Honorio, 2018; Chen et al., 2019) to sequentially identify leave nodes based on the estimated precision matrix. More recently, score-matching based methods for causal discovery have been proposed. This line of work typically assumes nonlinear additive Gaussian noise models (Rolland et al., 2022), while other settings are also considered (Montagna et al., 2023a,c,b; Sanchez et al., 2023).

B Proof of Proposition 1

Proposition 1 (Coupled single-node interventions)**.** *Consider a set of single-node interventions on the latent variables Z. The knowledge of which interventions share the same targets is equivalent to the knowledge of intervention targets up to variable permutation.*

Proof. The "⇐" side is trivial. We now prove the "⇒" side. For all single-node interventions $I^{(1)}, \ldots, I^{(m)}$, suppose that we know which of them share the same targets. That is, we can perform a partition of $\{I^{(1)}, \ldots, I^{(m)}\}$ into $\{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$, such that (1) the interventions in $\mathcal{I}_i$ share the same targets, and (2) the interventions in $\mathcal{I}_i$ do not share the same targets with the interventions in $\mathcal{I}_j, j \neq i$. Now suppose we perform a random permutation π of Z, denoted as Z_π, and assign $Z_{\pi(i)}$ as the target of interventions in $\mathcal{I}_i$. Clearly, one of the permutation π^* will lead to the correct assignment of intervention targets, but is unknown to us. This implies that we know the intervention targets (e.g., the assignment we perform with Z_π) up to variable permutation. □

C Proofs of Lemmas 1 and 2

C.1 Proof of Lemma 1

This lemma is implied by Rolland et al. (2022), and we provide the proof here for completeness.

Lemma 1. *Consider the data generating process in Eq.* (4) *and let Z_i be a sink node in DAG $\mathcal{G}_Z$. Then,*

$$\frac{\partial \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0 \quad \textit{for} \quad j \in [d].$$

Proof. Since the distribution $P_Z^{(u)}$ and the DAG $\mathcal{G}_Z$ satisfy the Markov property, we have

$$p^{(u)}(Z) = \prod_{k=1}^{n} p^{(u)}(Z_k \,|\, \mathrm{PA}(Z_k; \mathcal{G}_Z)).$$

By assumption, Z_i is a sink node in DAG $\mathcal{G}_Z$, which implies

$$\begin{aligned}\frac{\partial^2 \log p^{(u)}(Z)}{\partial Z_i^2} &= \frac{\partial^2 \log p^{(u)}(Z_i \,|\, \mathrm{PA}(Z_i; \mathcal{G}_Z))}{\partial Z_i^2} \\ &= \frac{\partial^2}{\partial Z_i^2}\left(-\frac{1}{2}\left(\frac{Z_i - f_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}{\sigma_i^{(u)}}\right)^2 - \frac{1}{2}\log(2\pi(\sigma_i^{(u)})^2)\right) \\ &= -\frac{1}{\sigma_i^{(u)}} \cdot \frac{\partial}{\partial Z_i}\left(\frac{Z_i - f_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}{\sigma_i^{(u)}}\right) \\ &= -\frac{1}{(\sigma_i^{(u)})^2},\end{aligned}$$

where $(\sigma_i^{(u)})^2$ is the variance of $\epsilon_i^{(u)}$. Therefore, for any Z_j, we have

$$\frac{\partial^3 \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0.$$

□

C.2 Proof of Lemma 2

Lemma 2. *Consider the data generating process in Eq.* (5) *and let Z_i be a sink node in DAG $\mathcal{G}_Z$. Then,*

$$\frac{\partial \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0 \quad \textit{for} \quad Z_j \notin \mathit{PA}(Z_i; \mathcal{G}_Z).$$

Proof. Since the distribution $P_Z^{(u)}$ and the DAG $\mathcal{G}_Z$ satisfy the Markov property, we have

$$p^{(u)}(Z) = \prod_{k=1}^{n} p^{(u)}(Z_k \,|\, \mathrm{PA}(Z_k; \mathcal{G}_Z)).$$

By assumption, Z_i is a sink node in DAG $\mathcal{G}_Z$, which implies

$$\begin{aligned}\frac{\partial^2 \log p^{(u)}(Z)}{\partial Z_i^2} &= \frac{\partial^2 \log p^{(u)}(Z_i \,|\, \mathrm{PA}(Z_i; \mathcal{G}_Z))}{\partial Z_i^2} \\ &= \frac{\partial^2}{\partial Z_i^2}\left(-\frac{1}{2}\left(\frac{Z_i - f_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}{\sigma_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}\right)^2\right) \\ &= -\frac{1}{\sigma_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))} \cdot \frac{\partial}{\partial Z_i}\left(\frac{Z_i - f_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}{\sigma_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z))}\right) \\ &= -\frac{1}{(\sigma_i^{(u)}(\mathrm{PA}(Z_i;\mathcal{G}_Z)))^2}.\end{aligned}$$

Therefore, for any $Z_j \notin \mathrm{PA}(Z_i; \mathcal{G}_Z)$, we have

$$\frac{\partial^3 \log p^{(u)}(Z)}{\partial Z_i^2 \partial Z_j} = 0.$$

□

D Proof of Theorem 1

Theorem 1 (Identifiability with latent ANMs)**.** *Consider the data generating process in Eqs. (1) and (4). Suppose that Assumptions 1 and 2, as well as the faithfulness assumption, hold. Let $\mathcal{G}_{\hat{Z}}$ and $\hat{Z}$ be the output of Algorithm 1. Then, there exists a permutation π of the estimated latent variables $\hat{Z}$, denoted as $\hat{Z}_\pi$, such that:*

1. *(Identifiability of $\mathcal{G}_Z$) $\mathcal{G}_{\hat{Z}_\pi}$ and $\mathcal{G}_Z$ are identical.*
2. *(Identifiability of Z) $\hat{Z}_{\pi(i)}$ is solely a function of a subset of $\{Z_i\} \cup IPA(Z_i; \mathcal{G}_Z)$.*

Proof. Let $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ be an output of Line 3 in Algorithm 1 during the $(n-1)$-th iteration, which corresponds to the output of Algorithm 1. Since both the true mixing function g and the estimated mixing function $\hat{g}$ are diffeomorphisms onto their images, the transformation from $\hat{Z}$ to Z is also a diffeomorphism. Therefore, there exists a permutation such that the Jacobian matrix $\frac{\partial Z_\alpha}{\partial \hat{Z}}$ has nonzero diagonal entries (e.g., see Zhang et al. (2024, Lemma 2)). Denote $Z_{\alpha[k]} \coloneqq (Z_{\alpha(l)})_{l=1}^k$. By the faithfulness assumption and Proposition 2 for the case of $(n-1)$-th iteration, we conclude that the DAGs $\mathcal{G}_{\hat{Z}}$ and $\mathcal{G}_{Z_\alpha}$ are identical, and that each latent variable $\hat{Z}_i, i \in [n]$ is solely a function of a subset of $Z_{\alpha(i)} \cup \bigcap_{k=i}^n \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$.

By definition, $\hat{Z}$ follows a causal order with respect to the latent DAG $\mathcal{G}_{\hat{Z}}$. Since $\mathcal{G}_{\hat{Z}}$ and $\mathcal{G}_{Z_\alpha}$ are identical, α is also a causal order of Z with respect to $\mathcal{G}_Z$. Under the faithfulness assumption, Proposition 3 implies that $\hat{Z}_i$ is solely a function of a subset of

$$Z_{\alpha(i)} \cup \bigcap_{k=i}^n \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}}) \subseteq Z_{\alpha(i)} \cup \mathrm{IPA}(Z_{\alpha(i)}; \mathcal{G}_Z).$$

By taking $\pi \coloneqq \alpha^{-1}$, we obtain the desired statements. □

D.1 Second-Order Partial Derivative of Latent Distribution

We provide the following lemma which will be used as the starting point to derive the third-order derivative in Eq. (7) in the proof of Proposition 2. This lemma is obtained from Zhang et al. (2024, Proposition 1). The proof involves change-of-variable formula (Ben-Israel, 1999; Gemici et al., 2016), chain rule, and property of Markov network (Lin, 1997), which is omitted here.

Lemma 3 (Second-order derivative)**.** *Consider the data generating process in Eq. (1). Suppose that we learn $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ to achieve Eq. (2). Then, we have*

$$\frac{\partial^2 \log p^{(u)}(\hat{Z})}{\partial \hat{Z}_k \partial \hat{Z}_l} = \sum_{i=1}^n \frac{\partial^2 \log p^{(u)}(Z)}{\partial Z_i^2} \frac{\partial Z_i}{\partial \hat{Z}_l} \frac{\partial Z_i}{\partial \hat{Z}_k} + \sum_{j=1}^n \sum_{i:\{Z_j, Z_i\} \in \mathcal{E}(\mathcal{M}_Z)} \frac{\partial^2 \log p^{(u)}(Z)}{\partial Z_i \partial Z_j} \frac{\partial Z_j}{\partial \hat{Z}_l} \frac{\partial Z_i}{\partial \hat{Z}_k}$$
$$+ \sum_{i=1}^n \frac{\partial \log p^{(u)}(Z)}{\partial Z_i} \frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l} + \frac{\partial^2 \log |\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l},$$

where $v \coloneqq g^{-1} \circ \hat{g}$ is a diffeomorphism.

D.2 Relation between Markov Network and Moral Graph

We provide the following lemma to relate the Markov network and moral graph, which is useful for proving Propositions 2 and 3.

Lemma 4 (Zhang et al. (2024, Lemma 1))**.** *Consider a DAG $\mathcal{G}_Z$ and distribution P_Z with its Markov Network $\mathcal{M}_Z$. Suppose that the Markov assumption holds. Then, the undirected graph defined by $\mathcal{M}_Z$ is a subgraph of the moral graph of DAG $\mathcal{G}_Z$.*

D.3 Proof of Lemma 5

The following lemma shows how diffeomorphism exists between subset of variables, which is useful for proving Proposition 2.

Lemma 5. *Suppose that the transformation from $\hat{Z} \in \mathbb{R}^n$ to $Z \in \mathbb{R}^n$ is a diffeomorphism, and that there exists $\mathcal{I} \subseteq [n]$ such that each $Z_i, i \in \mathcal{I}$ is not a function of $\hat{Z}_j, j \notin \mathcal{I}$. Then, the transformation from $\hat{Z}_{\mathcal{I}}$ to $Z_{\mathcal{I}}$ is a diffeomorphism.*

Proof. Let $J \coloneqq \frac{\partial Z}{\partial \hat{Z}}$ be the Jacobian matrix of the transformation from $\hat{Z}$ to Z. Since the transformation is a diffeomorphism, the Jacobian matrix J is invertible. By the Cayley-Hamilton theorem, for each value, its inverse $\frac{\partial \hat{Z}}{\partial Z} = J^{-1}$ can be written as $\sum_{k=0}^{n} c_k J^k$ for some coefficients $c_0, \ldots, c_n$. For $i \in \mathcal{I}$ and $j \notin \mathcal{I}$, since Z_i is, by definition, not a function of $\hat{Z}_j$, we clearly have $J_{i,j} = 0$. Also, it is straightforward to show $J_{i,:} J_{:,j} = 0$, which indicates $(J^2)_{i,j} = 0$. By mathematical induction, we obtain $(J^k)_{i,j} = 0$. This then implies

$$\frac{\partial \hat{Z}_i}{\partial Z_j} = \left(\frac{\partial \hat{Z}}{\partial Z}\right)_{i,j} = (J^{-1})_{i,j} = \left(\sum_{k=0}^{n} c_k J^k\right)_{i,j} = 0.$$

That is, each $\hat{Z}_i, i \in \mathcal{I}$ is not a function of $Z_j, j \notin \mathcal{I}$. This implies that each $Z_i, i \in \mathcal{I}$ is solely a function of $\hat{Z}_{\mathcal{I}}$, and vice versa, i.e., each $\hat{Z}_i, i \in \mathcal{I}$ is solely a function of $Z_{\mathcal{I}}$. Therefore, we conclude that the transformation from $\hat{Z}_{\mathcal{I}}$ to $Z_{\mathcal{I}}$ is a diffeomorphism. □

D.4 Proof of Proposition 2

We prove the following result via mathematical induction, which is crucial for the proof of Theorem 1. Specifically, it provides information on the identifiability of the latent DAG $\mathcal{G}_Z$ and latent variables Z in each iteration of Algorithm 1.

Proposition 2 (Output of t-th iteration)**.** *Consider the data generating process in Eqs. (1) and (4). Suppose that Assumptions 1 and 2, as well as the faithfulness assumption, hold. Let $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ be an output of Line 3 in Algorithm 1 during the t-th iteration. Let α be a permutation such that the Jacobian matrix $\frac{\partial Z_\alpha}{\partial \hat{Z}}$ has nonzero diagonal entries. Then, we have the following statements:*

(a) For $i \in [n], j = n+1-t, \ldots, n$, we have $\hat{Z}_i \to \hat{Z}_j$ in $\mathcal{G}_{\hat{Z}}$ if and only if $Z_{\alpha(i)} \to Z_{\alpha(j)}$ in $\mathcal{G}_Z$.

(b) Each latent variable $\hat{Z}_i, i \in [n]$ is solely a function of a subset of $Z_{\alpha(i)} \cup \bigcap_{k=\max(n+1-t,i)}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$.

(c) The transformation from $\hat{Z}_{[n-t]} \coloneqq (\hat{Z}_i)_{i=1}^{n-t}$ to $Z_{\alpha[n-t]} \coloneqq (Z_{\alpha(i)})_{i=1}^{n-t}$ is a diffeomorphism, and the Jacobian matrix $\frac{\partial Z_{\alpha[n-t]}}{\partial \hat{Z}_{[n-t]}}$ has nonzero diagonal entries.

Proof. We prove the proposition by mathematical induction from $t = 1$ to $n - 1$. We first provide the proof for the inductive step. Suppose that Proposition 2 holds for the case of t. In the following, we show that it also holds for the case of $t + 1$.

By definition, $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ denotes the output of Line 3 in Algorithm 1 during the $(t+1)$-th iteration, and α denotes a permutation such that the Jacobian matrix $\frac{\partial Z_\alpha}{\partial \hat{Z}}$ has nonzero diagonal entries, where $\hat{Z}$ is the estimated latent variables from Line 3 in Algorithm 1 during the $(t+1)$-th iteration. Due to the constraints (i) and (ii) in Line 2 of Algorithm 1, $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ is also a valid output of Line 3 in Algorithm 1 during the t-th iteration. (In this case, the same permutation α can also be used in the t-th iteration such that the corresponding Jacobian matrix $\frac{\partial Z_\alpha}{\partial \hat{Z}}$ has nonzero diagonal entries.) Therefore, by the induction hypothesis, Statements (a), (b), and (c) of Proposition 2 hold for $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ for the case of t, and we aim to further show that these statements also hold for $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$ for the case of $t + 1$.

By Statement (c) of Proposition 2 for the case of t, the transformation from $\hat{Z}_{[n-t]}$ to $Z_{\alpha[n-t]}$, denoted by v, is a diffeomorphism. By the change-of-variable formula, we have

$$\begin{aligned} p^{(u)}(\hat{Z}_{[n-t]}) |\det J_{v^{-1}}| &= p^{(u)}(Z_{\alpha[n-t]}) \\ \log p^{(u)}(\hat{Z}_{[n-t]}) &= \log p^{(u)}(Z_{\alpha[n-t]}) + \log |\det J_v|, \end{aligned} \tag{6}$$

where J_v denotes the Jacobian matrix of v. Suppose $\hat{Z}_l$ is a sink node in DAG $\mathcal{G}_{\hat{Z}_{[n-t]}}$. By Lemma 3, the second-order derivative w.r.t $\hat{Z}_l$ is given by

$$\begin{aligned}\frac{\partial^2 \log p^{(u)}(\hat{Z})}{\partial \hat{Z}_l^2} &= \sum_{i=t}^{n} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}^2} \left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l}\right)^2 + \sum_{j=t}^{n} \sum_{i:\{Z_{\alpha(i)},Z_{\alpha(j)}\}\in\mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)} \partial Z_{\alpha(j)}} \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \\ &= \sum_{i=t}^{n} \frac{\partial \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}} \frac{\partial^2 Z_{\alpha(i)}}{\partial \hat{Z}_l^2} + \frac{\partial^2 \log |\det J_v|}{\partial \hat{Z}_l^2}.\end{aligned}$$

Further taking third-order derivative w.r.t the sink node $\hat{Z}_l$ and applying Lemma 1, we have

$$\begin{aligned}0 = &\sum_{i:Z_{\alpha(i)}\in Z_{\alpha[n-t]}\setminus\mathcal{S}(\mathcal{G}_{Z_{\alpha[n-t]}})} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}^3} \left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l}\right)^3 \\ &+ \sum_{\substack{i,j: \\ \{Z_{\alpha(i)},Z_{\alpha(j)}\}\in\mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}}), \\ Z_{\alpha(i)}\notin\mathcal{S}(\mathcal{G}_{Z_{\alpha[n-t]}})}} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}^2 \partial Z_{\alpha(j)}} \left(2 \cdot \left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l}\right)^2 \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l}\right) \\ &+ \sum_{\substack{i,j,k: \\ i<j<k, \\ \{Z_{\alpha(i)},Z_{\alpha(j)}\},\{Z_{\alpha(j)},Z_{\alpha(k)}\},\{Z_{\alpha(i)},Z_{\alpha(k)}\}\in\mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})}} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)} \partial Z_{\alpha(j)} \partial Z_{\alpha(k)}} \left(6 \cdot \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \frac{\partial Z_{\alpha(k)}}{\partial \hat{Z}_l}\right) \\ &+ \sum_{i=1}^{d} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}^2} \left(3 \cdot \frac{\partial^2 Z_{\alpha(i)}}{\partial \hat{Z}_l^2} \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l}\right) \\ &+ \sum_{\substack{i,j: \\ i<j, \\ \{Z_{\alpha(i)},Z_{\alpha(j)}\}\in\mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})}} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)} \partial Z_{\alpha(j)}} \left(3 \cdot \frac{\partial^2 Z_{\alpha(j)}}{\partial \hat{Z}_l^2} \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} + 3 \cdot \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \frac{\partial^2 Z_{\alpha(i)}}{\partial \hat{Z}_l^2}\right) \\ &+ \sum_{i=1}^{n} \frac{\partial \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}} \frac{\partial^3 Z_{\alpha(i)}}{\partial \hat{Z}_l^3} + \frac{\partial^3 \log |\det J_v|}{\partial \hat{Z}_l^3}.\end{aligned} \tag{7}$$

By Assumption 2, there exist multiple values of u_j such that the above equation holds. Subtracting each equation corresponding to $u_j, j \neq 0$ with the equation corresponding to u_0, and using the assumption that the vectors formed by collecting the resulting coefficients (involving differences of partial derivatives) are linearly independent, we obtain

$$\left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l}\right)^3 = 0 \quad \Longleftrightarrow \quad \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} = 0 \quad \text{for} \quad Z_{\alpha(i)} \in Z_{\alpha[n-t]} \setminus \mathcal{S}(\mathcal{G}_{Z_{\alpha[n-t]}}). \tag{8}$$

Therefore, the non-sink nodes in $\mathcal{G}_{Z_{\alpha[n-t]}}$ cannot be a function of $\hat{Z}_l$ which is a sink node in in DAG $\mathcal{G}_{\hat{Z}_{[n-t]}}$. This implies that $Z_{\alpha(l)}$ is a sink node in DAG $\mathcal{G}_{Z_{\alpha[n-t]}}$, because otherwise $Z_{\alpha(l)}$ is not a function of $\hat{Z}_l$, which contradicts the induction hypothesis that the Jacobian matrix $\frac{\partial Z_{\alpha[n-t]}}{\partial \hat{Z}_{[n-t]}}$ has nonzero diagonal entries.

Recall that in Line 2 of Algorithm 1, the estimated Markov network $\mathcal{M}_{\hat{Z}_{[n-t]}}$ over $\hat{Z}_{[n-t]}$ has minimal number of edges. Thus, by Eq. (6), Assumption 1, and the induction hypothesis that the Jacobian matrix $\frac{\partial Z_{\alpha[n-t]}}{\partial \hat{Z}_{[n-t]}}$ has nonzero diagonal entries, Zhang et al. (2024, Theorem 2) implies that the Markov newtorks over $\hat{Z}_{[n-t]}$ and $Z_{\alpha[n-t]}$ are identical, i.e.,

$$\{\hat{Z}_i, \hat{Z}_j\} \in \mathcal{E}(\mathcal{M}_{\hat{Z}_{[n-t]}}) \quad \Longleftrightarrow \quad \{Z_{\alpha(i)}, Z_{\alpha(j)}\} \in \mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}}), \quad i, j \in [n-t], i \neq j. \tag{9}$$

The faithfulness assumption implies that the undirected graph defined by Markov network $\mathcal{M}_{Z_{\alpha[n-t]}}$ is the moral graph of DAG $\mathcal{G}_{Z_{\alpha[n-t]}}$ (e.g., see Zhang et al. (2024, Proposition 2)). Since $Z_{\alpha(l)}$ is a sink node in DAG $\mathcal{G}_{Z_{\alpha[n-t]}}$, each undirected edge $\{Z_{\alpha(i)}, Z_{\alpha(l)}\} \in \mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})$ in the moral graph implies a directed edge $Z_{\alpha(i)} \to Z_{\alpha(l)}$ in DAG $\mathcal{G}_Z$. Similar reasoning can be used to show that each undirected edge $\{\hat{Z}_i, \hat{Z}_l\} \in \mathcal{E}(\mathcal{M}_{\hat{Z}_{[n-t]}})$ in the moral graph implies a directed edge $\hat{Z}_i \to \hat{Z}_l$ in DAG $\mathcal{G}_{\hat{Z}}$. Therefore, we have

$$\hat{Z}_i \to \hat{Z}_l \text{ in } \mathcal{G}_{\hat{Z}} \quad \iff \quad Z_{\alpha(i)} \to Z_{\alpha(l)} \text{ in } \mathcal{G}_Z, \quad i \in [n].$$

In Line 3 of Algorithm 1, after reordering $\hat{Z}$ in $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$, we know that $\hat{Z}_{n-t}$ is a sink node in DAG $\mathcal{G}_{\hat{Z}_{[n-t]}}$. By the above relation, we then have

$$\hat{Z}_i \to \hat{Z}_{n-t} \text{ in } \mathcal{G}_{\hat{Z}} \quad \iff \quad Z_{\alpha(i)} \to Z_{\alpha(n-t)} \text{ in } \mathcal{G}_Z, \quad i \in [n]. \tag{10}$$

Recall that the induction hypothesis (i.e., Statement (a) of Proposition 2 for the case of t) indicates

$$\hat{Z}_i \to \hat{Z}_j \text{ in } \mathcal{G}_{\hat{Z}} \quad \iff \quad Z_{\alpha(i)} \to Z_{\alpha(j)} \text{ in } \mathcal{G}_Z, \quad i \in [n], j = n+1-t, \dots, n. \tag{11}$$

By Eqs. (10) and (11), we have shown that Statement (a) of Proposition 2 holds for the case of $t+1$.

Now we prove Statements (b) and (c) of Proposition 2 for the case of $t+1$. Recall that in Line 2 of Algorithm 1, the estimated Markov network $\mathcal{M}_{\hat{Z}_{[n-t]}}$ over $\hat{Z}_{[n-t]}$ has minimal number of edges. Thus, by Eq. (6), Assumption 1, and the induction hypothesis that the Jacobian matrix $\frac{\partial Z_{\alpha[n-t]}}{\partial \hat{Z}_{[n-t]}}$ has nonzero diagonal entries, Zhang et al. (2024, Proposition 4) and its proof imply the following statements:

- Statement(i): Each latent variable $Z_{\alpha(i)}, i \in [n-t]$ is solely a function of a subset of $\hat{Z}_i \cup \Psi(\hat{Z}_i; \mathcal{M}_{\hat{Z}_{[n-t]}})$.
- Statement(ii): Each latent variable $\hat{Z}_i, i \in [n-t]$ is solely a function of a subset of $Z_{\alpha(i)} \cup \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[n-t]}})$.

By the induction hypothesis (i.e., Statement (b) of Proposition 2 for the case of t), each latent variable $\hat{Z}_i, i \in [n]$ is solely a function of a subset of $Z_{\alpha(i)} \cup \bigcap_{k=\max(n+1-t,i)}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$. By Statement (ii), one can straightforwardly show that each latent variable $\hat{Z}_i, i \in [n]$ is solely a function of a subset of $Z_{\alpha(i)} \cup \bigcap_{k=\max(n-t,i)}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$, indicating that Statement (b) of Proposition 2 hold for the case of $t+1$.

In Line 3 of Algorithm 1, after reordering $\hat{Z}$ in $(\hat{g}, p_{\hat{Z}}, \mathcal{G}_{\hat{Z}})$, we know that $\hat{Z}_{n-t}$ is a sink node in DAG $\mathcal{G}_{\hat{Z}_{[n-t]}}$. First, by Eq. (8), the non-sink nodes in $\mathcal{G}_{Z_{\alpha[n-t]}}$ cannot be a function of $\hat{Z}_{n-t}$. This also indicates that $Z_{\alpha(n-t)}$ is a sink node in $\mathcal{G}_{Z_{\alpha[n-t]}}$, due to the induction hypothesis that the Jacobian matrix $\frac{\partial Z_{\alpha[n-t]}}{\partial \hat{Z}_{[n-t]}}$ has nonzero diagonal entries. Second, suppose that $Z_{\alpha(k)}$ is a sink node in $\mathcal{G}_{Z_{\alpha[n-t]}}$, and that it is distinct from $Z_{\alpha(n-t)}$. Since both $Z_{\alpha(n-t)}$ and $Z_{\alpha(k)}$ are sink nodes in DAG $\mathcal{G}_{Z_{\alpha[n-t]}}$, clearly they cannot be adjacent to each other or share a common child in DAG $\mathcal{G}_{Z_{\alpha[n-t]}}$. By Lemma 4, $Z_{\alpha(n-t)}$ and $Z_{\alpha(k)}$ cannot be adjacent in Markov network $\mathcal{M}_{Z_{\alpha[n-t]}}$. By Eq. (9), $\hat{Z}_{n-t}$ and $\hat{Z}_k$ cannot be adjacent in Markov network $\mathcal{M}_{\hat{Z}_{[n-t]}}$. This implies $\hat{Z}_{n-t} \notin \Psi(\hat{Z}_k; \mathcal{M}_{\hat{Z}_{[n-t]}})$, which, by Statement (i) derived above, indicates that $Z_{\alpha(k)}$ cannot be a function of $\hat{Z}_{n-t}$. Combining both cases (for sink nodes and non-sink nodes), we conclude that all variables in $Z_{\alpha[n-t]}$, except $Z_{\alpha(n-t)}$, cannot be a function of $\hat{Z}_{n-t}$. Therefore, by the induction hypothesis about diffeomorphism from $\hat{Z}_{[n-t]}$ to $Z_{\alpha[n-t]}$, applying Lemma 5 implies that the transformation from $\hat{Z}_{[n-1-t]}$ to $Z_{\alpha[n-1-t]}$ is a diffeomorphism. Also, clearly the Jacobian matrix $\frac{\partial Z_{\alpha[n-1-t]}}{\partial \hat{Z}_{[n-1-t]}}$ has nonzero diagonal entries. That is, we have shown that Statement (c) of Proposition 2 holds for the case of $t+1$.

Up until now, we have shown that the inductive step holds for the proof of Proposition 2. The same technique applies to the base case of $t = 1$, which is omitted here. Specifically, for the base case, we rely on the assumption that both the true mixing function g and the estimated mixing function $\hat{g}$ are diffeomorphisms onto their images, indicating that there the transformation from $\hat{Z}$ to Z is a diffeomorphism. $\square$

D.5 Proof of Proposition 3

The following result is used in the proof of Theorem 1, specifically for the identifiability of latent variables Z. It relates the intimate neighbors in different Markov networks to the intimate parents in the latent DAG.

Proposition 3. *Consider the data generating process in Eq.* (1)*. Let α be a causal order of variables $Z = (Z_1, \dots, Z_n)$ with respect to DAG $\mathcal{G}_Z$ and denote $Z_{\alpha[k]} \coloneqq (Z_{\alpha(l)})_{l=1}^{k}$. Under the faithfulness assumption, we have*

$$\bigcap_{k=i}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}}) \subseteq \mathit{IPA}(Z_{\alpha(i)}; \mathcal{G}_Z) \quad \text{for} \quad i \in [n-1].$$

Proof. It suffices to show that $Z_{\alpha(j)} \in \bigcap_{k=i}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$ implies $Z_{\alpha(j)} \in \mathrm{IPA}(Z_{\alpha(i)}; \mathcal{G}_Z)$ for $i \in [n-1]$. We provide a proof by contrapositive, i.e., we aim to show that $Z_{\alpha(j)} \notin \mathrm{IPA}(Z_{\alpha(i)}; \mathcal{G}_Z)$ implies

$$Z_{\alpha(j)} \notin \bigcap_{k=i}^{n} \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}}). \tag{12}$$

Now suppose $Z_{\alpha(j)} \notin \mathrm{IPA}(Z_{\alpha(i)}; \mathcal{G}_Z)$. By the definition of intimate parents, we have the following cases.

Case 1: $Z_{\alpha(j)}$ is not a parent of $Z_{\alpha(i)}$. We need to consider the following cases.

- **Case 1(a):** $Z_{\alpha(j)}$ is a child of $Z_{\alpha(i)}$. This means $i < j$. For $i \leq k < j$, we have $Z_{\alpha(j)} \notin Z_{\alpha[k]}$ and thus $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$, which implies Eq. (12).
- **Case 1(b):** $Z_{\alpha(j)}$ is a spouse of $Z_{\alpha(i)}$. By definition of spouse, $Z_{\alpha(j)}$ is not adjacent to $Z_{\alpha(i)}$ in $\mathcal{G}_Z$. If $i < j$, then for $i \leq k < j$, we have $Z_{\alpha(j)} \notin Z_{\alpha[k]}$ and thus $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[k]}})$, which implies Eq. (12). Otherwise, we have $i > j$. In this case, consider the latent DAG $\mathcal{G}_{Z_{\alpha[i]}}$, where $Z_{\alpha(i)}$ is a sink node. Clearly, $Z_{\alpha(j)}$ is not a parent, child, or spouse of $Z_{\alpha(i)}$ in DAG $\mathcal{G}_{Z_{\alpha[i]}}$. Therefore, $Z_{\alpha(j)}$ and $Z_{\alpha(i)}$ are not adjacent in the moral graph of $\mathcal{G}_{Z_{\alpha[i]}}$, which, by Lemma 4, indicates that they are not adjacent in the Markov network $\mathcal{M}_{Z_{\alpha[i]}}$. This implies $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[i]}})$ and thus Eq. (12).
- **Case 1(c):** $Z_{\alpha(j)}$ is not a child or spouse of $Z_{\alpha(i)}$. In this case, $Z_{\alpha(j)}$ is not a parent, child, or spouse of $Z_{\alpha(i)}$. Therefore, $Z_{\alpha(j)}$ and $Z_{\alpha(i)}$ are not adjacent in the moral graph of $\mathcal{G}_Z$, which, by Lemma 4, indicates that they are not adjacent in the Markov network $\mathcal{M}_Z$. This implies $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_Z)$ and thus Eq. (12).

Case 2: $Z_{\alpha(j)}$ is not a parent of some child of $Z_{\alpha(i)}$, denoted as $Z_{\alpha(l)}$. This means $l > i$. In this case, consider the latent DAG $\mathcal{G}_{Z_{\alpha[l]}}$, where $Z_{\alpha(l)}$ is a sink node. Clearly, $Z_{\alpha(j)}$ is not a parent, child, or spouse of $Z_{\alpha(l)}$ in DAG $\mathcal{G}_{Z_{\alpha[l]}}$. Therefore, $Z_{\alpha(j)}$ and $Z_{\alpha(l)}$ are not adjacent in the moral graph of $\mathcal{G}_{Z_{\alpha[l]}}$. By Lemma 4, this indicates that $Z_{\alpha(j)}$ is not adjacent to $Z_{\alpha(l)}$ in the Markov network $\mathcal{M}_{Z_{\alpha[l]}}$, which, under the faithfulness assumption, is a neighbor of $Z_{\alpha(i)}$ in the Markov network $\mathcal{M}_{Z_{\alpha[l]}}$. This implies $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_{Z_{\alpha[l]}})$ and thus Eq. (12).

Case 3: $Z_{\alpha(j)}$ is not adjacent to, and not a spouse of, some spouse of $Z_{\alpha(i)}$, denoted as $Z_{\alpha(l)}$. In this case, $Z_{\alpha(j)}$ is not a parent, child, or spouse of $Z_{\alpha(l)}$. Therefore, $Z_{\alpha(j)}$ and $Z_{\alpha(l)}$ are not adjacent in the moral graph of $\mathcal{G}_Z$. By Lemma 4, this indicates that $Z_{\alpha(j)}$ is not adjacent to $Z_{\alpha(l)}$ in the Markov network $\mathcal{M}_Z$, which, under the faithfulness assumption, is a neighbor of $Z_{\alpha(i)}$ in the Markov network $\mathcal{M}_Z$. This implies $Z_{\alpha(j)} \notin \Psi(Z_{\alpha(i)}; \mathcal{M}_Z)$ and thus Eq. (12). □

E Proof of Theorem 2

Theorem 2 (Identifiability with latent HNMs)**.** *Consider the data generating process in Eqs.* (1) *and* (5)*. Suppose that Assumptions 1 and 3, as well as the faithfulness assumption, hold. Let $\mathcal{G}_{\hat{Z}}$ and $\hat{Z}$ be the output of Algorithm 1. Then, there exists a permutation π of the estimated latent variables $\hat{Z}$, denoted as $\hat{Z}_\pi$, such that:*

1. *(Identifiability of $\mathcal{G}_Z$) $\mathcal{G}_{\hat{Z}_\pi}$ and $\mathcal{G}_Z$ are identical.*
2. *(Identifiability of Z) $\hat{Z}_{\pi(i)}$ is solely a function of a subset of $\{Z_i\} \cup \mathit{IPA}(Z_i; \mathcal{G}_Z)$.*

Proof. The proof here is identical to that of Theorem 1, which leverages Propositions 2 and 3. Therefore, we omit the proof here. The only (minor) difference is that the third-order derivative in Eq. (7) is instead given by:

$$
\begin{aligned}
0 = & \sum_{i: Z_{\alpha(i)} \in Z_{\alpha[n-t]} \setminus \mathcal{S}(\mathcal{G}_{Z_{\alpha[n-t]}})} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z^3_{\alpha(i)}} \left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \right)^3 \\
& + \sum_{\substack{i,j: \\ \{Z_{\alpha(i)}, Z_{\alpha(j)}\} \in \mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})}} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z^2_{\alpha(i)} \partial Z_{\alpha(j)}} \left(2 \cdot \left(\frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \right)^2 \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \right) \\
& + \sum_{\substack{i,j,k: \\ i<j<k, \\ \{Z_{\alpha(i)}, Z_{\alpha(j)}\}, \{Z_{\alpha(j)}, Z_{\alpha(k)}\}, \{Z_{\alpha(i)}, Z_{\alpha(k)}\} \in \mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})}} \frac{\partial^3 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)} \partial Z_{\alpha(j)} \partial Z_{\alpha(k)}} \left(6 \cdot \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \frac{\partial Z_{\alpha(k)}}{\partial \hat{Z}_l} \right) \\
& + \sum_{i=1}^{d} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z^2_{\alpha(i)}} \left(3 \cdot \frac{\partial^2 Z_{\alpha(i)}}{\partial \hat{Z}^2_l} \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} \right) \\
& + \sum_{\substack{i,j: \\ i<j, \\ \{Z_{\alpha(i)}, Z_{\alpha(j)}\} \in \mathcal{E}(\mathcal{M}_{Z_{\alpha[n-t]}})}} \frac{\partial^2 \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)} \partial Z_{\alpha(j)}} \left(3 \cdot \frac{\partial^2 Z_{\alpha(j)}}{\partial \hat{Z}^2_l} \frac{\partial Z_{\alpha(i)}}{\partial \hat{Z}_l} + 3 \cdot \frac{\partial Z_{\alpha(j)}}{\partial \hat{Z}_l} \frac{\partial^2 Z_{\alpha(i)}}{\partial \hat{Z}^2_l} \right) \\
& + \sum_{i=1}^{n} \frac{\partial \log p^{(u)}(Z_{\alpha[n-t]})}{\partial Z_{\alpha(i)}} \frac{\partial^3 Z_{\alpha(i)}}{\partial \hat{Z}^3_l} + \frac{\partial^3 \log |\det J_v|}{\partial \hat{Z}^3_l}.
\end{aligned}
$$

□

F Further Empirical Studies

F.1 Selecting Number of Latent Variables

We provide empirical studies to demonstrate how to determine the number of latent variables. Specifically, according to Fig. 3, one can in practice perform model selection to select the number of latent variables based on the evidence lower bound (ELBO) loss.

F.2 Selecting Hyperparameters

We discuss how to select the hyperparameters for sparsity. Here, we choose λ_1 based on the ELBO loss and provide an example in Fig. 4. If the structure is overly sparse, it fails to reconstruct the input data and the reconstruction error will become high. If the structure is overly dense, it involves many unwanted edges and the KL divergence will become high.

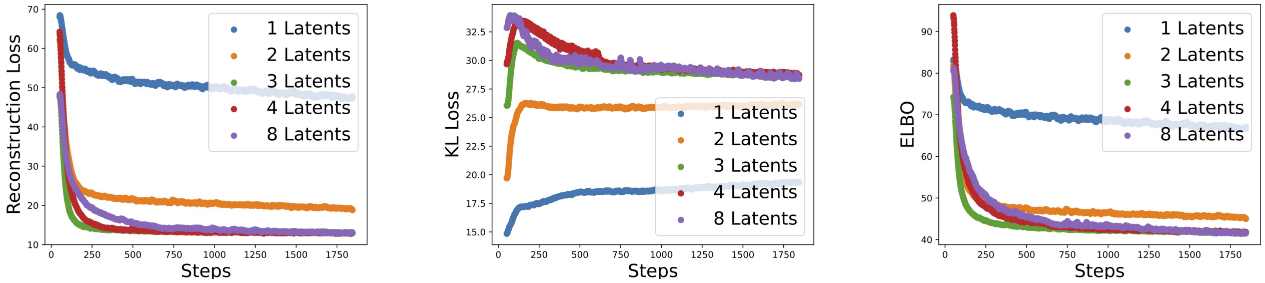

Figure 3: The losses versus the training steps with different number of estimated latent variables. The ground truth number of latent variables is 3. When we set the estimated number of latent variables to be low, e.g., 1 or 2, the reconstruction is poor, showing that they are unable to match the distributions. This suggests that the number of estimated latent variables can be selected based on the ELBO loss.

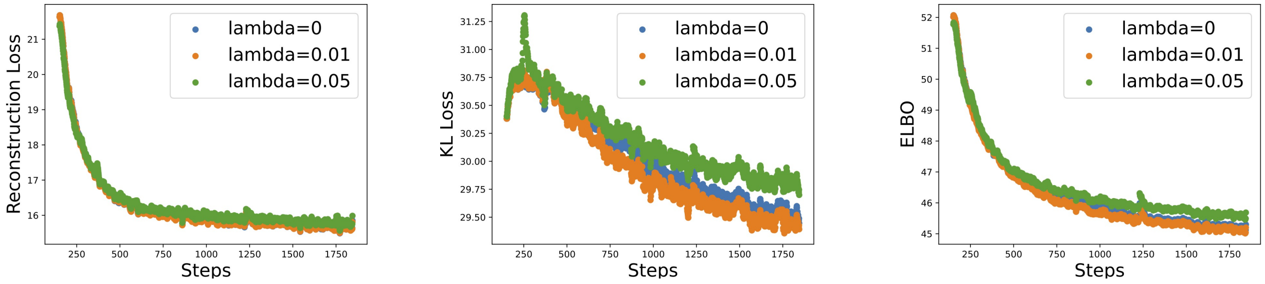

Figure 4: The losses versus the training steps with different values of λ_1. If we set $\lambda_1 = 0.05$, which may be too high for the data, the ELBO loss is worst. If we set $\lambda_1 = 0$, the estimated graph is dense, i.e., estimated variables are unnecessarily dependent. When we set $\lambda_1 = 0.01$, we achieve an ELBO loss that is as good as the ELBO loss with $\lambda_1 = 0$, but with fewer edges.