Causal Representation Learning from Multiple Distributions: A General Setting

Kun Zhang [*1,2] **Shaoan Xie** [*1] **Ignavier Ng** [*1] **Yujia Zheng** [1]

Abstract

In many problems, the measured variables (e.g., image pixels) are just mathematical functions of the latent causal variables (e.g., the underlying concepts or objects). For the purpose of making predictions in changing environments or making proper changes to the system, it is helpful to recover the latent causal variables Z_i and their causal relations represented by graph $\mathcal{G}_Z$. This problem has recently been known as causal representation learning. This paper is concerned with a general, completely nonparametric setting of causal representation learning from multiple distributions (arising from heterogeneous data or nonstationary time series), without assuming hard interventions behind distribution changes. We aim to develop general solutions in this fundamental case; as a by product, this helps see the unique benefit offered by other assumptions such as parametric causal models or hard interventions. We show that under the sparsity constraint on the recovered graph over the latent variables and suitable sufficient change conditions on the causal influences, interestingly, one can recover the moralized graph of the underlying directed acyclic graph, and the recovered latent variables and their relations are related to the underlying causal model in a specific, nontrivial way. In some cases, most latent variables can even be recovered up to component-wise transformations. Experimental results verify our theoretical claims.

1. Introduction

Causal representation learning holds paramount significance across numerous fields, offering insights into intricate relationships within datasets. Most traditional methodologies (e.g., causal discovery) assume the observation of causal variables. This assumption, however reasonable, falls short in complex scenarios involving indirect measurements, such as electronic signals, image pixels, and linguistic tokens. Moreover, there are usually changes on the causal mechanisms in real-world, such as the heterogeneous or nonstationary data. Identifying the latent causal variables and their structures together with the change of the causal mechanism is in pressing need to understand the complicated real-world causal process. This has been recently known as causal representation learning (Schölkopf et al., 2021).

It is worth noting that identifying only the latent causal variables but not the structure among them, is already a considerable challenge. In the i.i.d. case, different latent representations can explain the same observations equally well, while not all of them are consistent with the true causal process. For instance, nonlinear independent component analysis (ICA), where a set of observed variables X is represented as a mixture of independent latent variables Z, i.e, $X = g(Z)$, is known to be unidentifiable without additional assumptions (Comon, 1994). While being a strictly easier task since there are no relations among latent variables, the identifiability of nonlinear ICA often relies on conditions on distributional assumptions (non-i.i.d. data) (Hyvärinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a; Sorrenson et al., 2020; Lachapelle et al., 2022; Hälvä & Hyvärinen, 2020; Hälvä et al., 2021; Yao et al., 2022) or specific functional constraints (Comon, 1994; Hyvärinen & Pajunen, 1999; Taleb & Jutten, 1999; Buchholz et al., 2022; Zheng et al., 2022; Zheng & Zhang, 2023).

To generalize beyond the independent latent variables and achieve causal representation learning (recovering the latent variables and their causal structure), recent advances either introduce additional experiments in the forms of interventional or counterfactual data, or place more restrictive parametric or graphical assumptions on the latent causal model. For observational data, various graphical conditions have been proposed together with parametric assumptions such as linearity (Silva et al., 2006; Cai et al., 2019; Xie et al., 2020; 2022; Adams et al., 2021; Huang et al., 2022) and discreteness (Kivva et al., 2021). For interventional data, single-node interventions have been considered together with parametric assumptions (e.g., linearity) on the mixing function (Varici et al., 2023; Ahuja et al., 2023; Buchholz et al., 2022) or also on the latent

[*]Equal contribution [1]Carnegie Mellon University [2]Mohamed bin Zayed University of Artificial Intelligence.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

causal model (Squires et al., 2023). The nonparametric settings for both the mixing function and causal model have been explored by (Brehmer et al., 2022; von Kügelgen et al., 2023; Jiang & Aragam, 2023) together with additional assumptions on counterfactual views (Brehmer et al., 2022), distinct paired interventions (von Kügelgen et al., 2023), and graphical conditions (Jiang & Aragam, 2023).

Despite the exciting developments in the field, one fundamental question pertinent to causal representation learning from multiple distributions remains unanswered–in the most general situation, without assuming parametric models on the data-generating process or the existence of hard interventions in the data, what information of the latent variables and the latent structure can be recovered? This paper attempts to provide an answer to it, which, surprisingly, shows that each latent variable can be recovered up to clearly defined indeterminacies. It suggests what we can achieve in the general case and furthermore, what unique contribution the typical assumptions that are currently made in causal representation learning from multiple distributions make towards complete identifiability of the latent variables (up to component-wise transformations). This may make it possible to figure out what minimal assumptions are needed to achieve complete identifiability, given partial knowledge of the system.

Contributions. Concretely, as our contributions, we show that under the sparsity constraint on the recovered graph over the latent variables and suitable sufficient change conditions on the causal influences, interestingly, one can recover the moralized graph of the underlying directed acyclic graph (Theorem 2), and the recovered latent variables and their relations are related to the underlying causal model in a specific, nontrivial way (Theorem 3)–each latent variables is recovered as a function of itself and its so-called *intimate neighbors* in the Markov network implied by the true causal structure over the latent variables. Depending on the properties of the true causal structure over latent variables, the set of intimate neighbors might even be empty, in which case the corresponding latent variables can be recovered up to component-wise transformations (Remark 1). Lastly, we show how the recovered moralized graph relates to the underlying causal graph under new relaxations of faithfulness assumption (Proposition 2). Simulation studies verified our theoretical findings.

2. Problem Setting

Let $X = (X_1, \dots, X_d)$ be a d-dimensional vector that represents the observed variables (e.g., image pixels). We assume that they are generated by n latent causal variables $Z = (Z_1, \dots, Z_n)$ via a nonlinear mixing function $g : \mathbb{R}^n \to \mathbb{R}^d$ ($d \geq n$), which is a $\mathcal{C}^2$-diffeomorphism onto its image $\mathcal{X} \subseteq \mathbb{R}^d$. Furthermore, the variables Z_i's are assumed to follow a structural equation model (SEM) (Pearl, 2000). Putting them together, the underlying data generating process can be written as

$$\underbrace{X = g(Z)}_{\text{Nonlinear mixing}}, \quad \underbrace{Z_i = f_i(\text{PA}(Z_i), \epsilon_i; \theta_i), i = 1, \dots, n}_{\text{Latent SEM}}. \tag{1}$$

where $\text{PA}(Z_i)$ denotes the parents of variable Z_i, ϵ_i's are exogenous noise variables that are mutually independent, and θ_i denotes the latent (changing) factor (or effective parameters) associated with each model. Here, the data generating process of each latent variable Z_i may change, e.g., across domains or over time, governed by the corresponding latent factor θ_i; it is commonplace to encounter such changes in causal mechanisms in practice (arising from heterogeneous data or nonstationary time series). In addition, interventional data can be seen as a special type of change, which qualitatively restructure the causal relations. As their names suggest, we assume that the variables X are observed, while the latent causal variables Z and latent factors $\theta = (\theta_1, \dots, \theta_n)$ are unobserved.

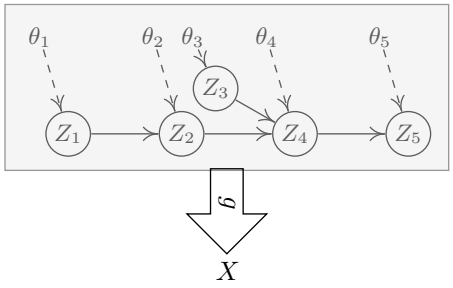


Figure 1: The generating process for each latent causal variable Z_i changes, governed by a latent factor θ_i. The observed variables X are generated by $X = g(Z)$ with a nonlinear mixing function g.

Let $P_{X;\theta}$ and $P_{Z;\theta}$ be the distributions of X and Z, respectively, and their probability density functions be $p_X(X;\theta)$ and $p_Z(Z;\theta)$, respectively.[1] To lighten the notation, we drop the subscript in the density when the context is clear. The latent SEM in Eq. (1) induces a causal graph $\mathcal{G}_Z$ with vertices $\{Z_i\}_{i=1}^n$ and edges $Z_j \to Z_i$ if and only if $Z_j \in \text{PA}(Z_i)$. We assume that $\mathcal{G}_Z$ is acyclic, i.e., a directed acyclic graph (DAG). This implies that the distribution of variables Z satisfy the Markov property w.r.t. DAG $\mathcal{G}_Z$ (Pearl, 2000), i.e., $p(Z;\theta) = \prod_{i=1}^n p(Z_i \,|\, \text{PA}(Z_i);\theta_i)$. We provide an example of the data generating process in Eq. (1) and its corresponding latent DAG $\mathcal{G}_Z$ in Figure 1. Given samples of the observed variables X arising from multiple distributions (or domains), say $\Theta = \{\theta^{(1)}, \dots, \theta^{(m)}\}$, our goal is to recover the latent causal variables Z and their causal relations up to minor indeterminacies.

[1] With a slight abuse of notation, we use the same capital letters X and Z to denote the variables and their values when the context is clear.

3. Learning Causal Representations from Multiple Distributions

In this section, we provide theoretical results to show how one is able to recover the underlying latent causal variables and their causal relations up to certain indeterminacies from multiple distributions. Specifically, we show that under sparsity constraint on the recovered graph over the latent variables and suitable sufficient change conditions on the causal influences, the recovered latent variables are related to the true ones in a specific, nontrivial way. Such results serve as the foundation of our algorithm in Section 4.

To start with, we estimate a model $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ that assumes the same data generating process as in Eq. (1) and matches the true distribution of X in different domains:

$$p_X(X'; \theta^{(u)}) = p_{\hat{X}}(X'; \hat{\theta}^{(u)}), \forall\, \theta^{(u)} \in \Theta,\ X' \in \mathcal{X}^{(u)}, \quad (2)$$

where $\theta^{(u)}$ denotes the latent factor in the u-th domain, and $\mathcal{X}^{(u)}$ is the image of function g in the u-th domain. Here, X and $\hat{X}$ are generated by the true model (g, f, p_Z, Θ) and the estimated model $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$, respectively.

A key ingredient of our results is the Markov network that represents conditional dependencies among random variables via an undirected graph. Let $\mathcal{M}_Z$ be the Markov network over variables Z, i.e., with vertices $\{Z_i\}_{i=1}^n$ and edges $\{Z_i, Z_j\} \in \mathcal{E}(\mathcal{M}_Z)$ if and only if $Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}}$.[2] Also, we denote by $|\mathcal{M}_Z|$ the number of undirected edges in the Markov network. In Section 3.1, apart from showing how to estimate the underlying latent causal variables up to certain indeterminacies, we also show that such latent Markov network $\mathcal{M}_Z$ can be recovered up to isomorphism. To achieve so, we make use of the following property (assuming that p_Z is twice differentiable):

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

Such a connection between pairwise conditional independence and cross derivatives of the density function has been noted by Lin (1997) and utilized in Markov network learning for observed variables (Zheng et al., 2023). With the recovered latent Markov network structure, we provide results in Section 3.2 to show how it relates to the moralized graph of true latent causal DAG $\mathcal{G}_Z$, by exploiting a specific type of faithfulness assumption that is considerably weaker than the standard faithfulness assumption used in the literature of causal discovery (Spirtes et al., 2001).

3.1. Recovering Latent Causal Variables and Latent Markov Network

We consider a general, completely nonparametric setting of causal representation learning from multiple distributions. Specifically, we show how one can recover the latent causal variables and the Markov network structure among them up to minor indeterminacies, by leveraging sparsity constraint and sufficient change conditions on the causal mechanisms. Notably, in some cases, most latent variables can even be recovered up to component-wise transformations.

We start with the following result that provides information about the derivative of true latent causal variables Z with respect to the estimated ones $\hat{Z}$, according to their corresponding Markov networks $\mathcal{M}_Z$ and $\mathcal{M}_{\hat{Z}}$. Result of this form is often used in the proof of nonlinear ICA to obtain identifiability of component-wise nonlinear transformations (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019). At the same time, our result here is different from that of nonlinear ICA as it allows for causal relations among latent variables. This result serves as the backbone of our further identifiability results in this section.

Proposition 1. *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose the following assumptions hold:*

- *A1 (Smooth and positive density): The probability density function of latent causal variables, i.e., p_Z, is twice continuously differentiable and positive in $\mathbb{R}^n$.*
- *A2 (Sufficient changes): For each value of Z, there exist $2n + |\mathcal{M}_Z| + 1$ values of θ, i.e., $\theta^{(u)}$ with $u = 0, \ldots, 2n+|\mathcal{M}_Z|$, such that the vectors $w(Z,u) - w(z,0)$ with $u = 1, \ldots, 2n + |\mathcal{M}_Z|$ are linearly independent, where vector $w(Z,u)$ is defined as follows:*[3]

$$w(Z,u) = \left(\frac{\partial \log p(Z;\theta^{(u)})}{\partial Z_i}\right)_{i\in[n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i^2}\right)_{i\in[n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i \partial Z_j}\right)_{\{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_Z),\, i<j}.$$

Suppose that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2). Then, for every pair of estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are **not adjacent in the Markov network** *$\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$, we have the following statements:*

(a) For each true latent causal variable Z_i, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_i}{\partial \hat{Z}_l} = 0. \quad (4)$$

(b) For each pair of true latent causal variables Z_i and Z_j that are adjacent in the Markov network $\mathcal{M}_Z$, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_j}{\partial \hat{Z}_l} = 0. \quad (5)$$

[2] We use $[n]$ to denote $\{1, \ldots, n\}$ and $Z_{[n]\setminus\{i,j\}}$ to denote $\{Z_i\}_{i=1}^n \setminus \{Z_i, Z_j\}$.

[3] We denote by $\oplus$ the vector concatenation symbol. Also, the order in the mixed partial derivatives can be interchanged.

The proof is provided in Appendix B, which leverages the property of Markov network in Eq. (3). Assumption A2 can be viewed as suitable sufficient change conditions on the causal influences across different domains. It is worth noting that the requirement of a sufficient number of domains has been commonly adopted in the literature (e.g., see Hyvärinen et al. (2023) for a recent survey), such as visual disentanglement (Khemakhem et al., 2020b), domain adaptation (Kong et al., 2022), video analysis (Yao et al., 2021), and image-to-image translation (Xie et al., 2022). Also, we do not specify exactly how to learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2), and leave the door open for different approaches to be used, such as normalizing flow and variational approaches. For example, we adopt a variational approach in Section 4.

In Theorem 1, Eqs. (4) and (5) hold for every sample of Z. Intuitively, one may expect that Eq. (4) implies either $\frac{\partial Z_i}{\partial \hat{Z}_k} = 0$ for all samples of Z, or $\frac{\partial Z_i}{\partial \hat{Z}_l} = 0$ for all samples, i.e., the zero entries in the Jacobian matrix (of the function from $\hat{Z}$ to Z) remain in the same positions across different samples. If this conclusion holds true, it indicates that the true latent variable Z_i cannot be a function of both estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$, which is helpful for disentanglement. The same reasoning applies to Eq. (5). In fact, similar conclusion can often be obtained in the proof of identifiability for nonlinear ICA (Hyvärinen et al., 2019), by leveraging the continuity and invertibility of the Jacobian matrix.

However, this conclusion in general does not hold in our setting (that allows for causal relations among latent variables Z) without any constraint on the sparsity of recovered Markov network, for which counterexamples exist. The reason is that each of Eqs. (4) and (5) correspond to a pair of recovered latent variables $\hat{Z}$ that are not adjacent in the Markov network $\mathcal{M}_{\hat{Z}}$, and can be viewed as a specific form of restriction on the Jacobian matrix (of the function from $\hat{Z}$ to Z). When the recovered Markov network is relatively dense, less restrictions are imposed on the Jacobian matrix, and thus there are possibilities for the aforementioned zero entries to switch positions across different samples. Interestingly, incorporating sparsity constraint on the recovered Markov network during estimation can help eliminate these possibilities, formally described below.

Theorem 1 (Relations among true and recovered latent causal variables)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, for every pair of estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are* **not adjacent in the Markov network** *$\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$, we have the following statements:*

(a) Each true latent causal variable Z_i is a function of at most one of $\hat{Z}_k$ and $\hat{Z}_l$.

(b) For each pair of true latent causal variables Z_i and Z_j that are adjacent in the Markov network $\mathcal{M}_Z$ over Z, at most one of them is a function of $\hat{Z}_k$ or $\hat{Z}_l$.

The proof can be found in Appendix D. The above result sheds light on how each pair of the estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are not adjacent in Markov network $\mathcal{M}_{\hat{Z}}$ relate to the true latent causal variables Z, thus providing information for further disentanglement. Furthermore, note that a trivial solution would be a complete graph over $\hat{Z}$ without any constraint on the estimating process. Apart from providing information for disentanglement, we show below that incorporating sparsity constraint on the recovered Markov network also helps avoid this trivial solution and recover the underlying Markov network up to isomorphism. The proof is given in Appendix C.

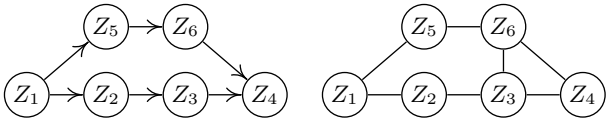
Theorem 2 (Identifiability of latent Markov network)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, the recovered latent Markov network $\mathcal{M}_{\hat{Z}}$ is isomorphic to the true latent Markov network $\mathcal{M}_Z$.*

In addition to recovering the underlying Markov network $\mathcal{M}_Z$, we show that the sparsity constraint on the recovered Markov network $\mathcal{M}_{\hat{Z}}$ also allows us to recover the underlying latent causal variables Z up to specific, relatively minor indeterminacies. In the result, the following variable set, termed *intimate neighbor set*, plays an important role:

$$\Psi_{Z_i} := \{Z_j \mid Z_j, j \neq i, \text{ is adjacent to } Z_i \text{ and all other neighbors of } Z_i \text{ in } \mathcal{M}_Z\}.$$

For example, according to the Markov network implied by $\mathcal{G}_Z$ in Figure 1, $\Psi_{Z_1} = \{Z_2\}$, $\Psi_{Z_2} = \varnothing$, where $\varnothing$ denotes the empty set, $\Psi_{Z_3} = \{Z_2, Z_4\}$, $\Psi_{Z_4} = \varnothing$, and $\Psi_{Z_5} = \{Z_4\}$. As another example, according to the Markov network in Figure 2(b), which is implied by the DAG in Figure 2(a), we have $\Psi_{Z_i} = \varnothing$ for $i = 1, 2, 3, 5, 6$ and $\Psi_{Z_4} = \{Z_3, Z_6\}$.

Theorem 3 (Identifiability of latent causal variables)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Let N_{Z_i} be the set of neighbors of variable Z_i in $\mathcal{M}_Z$. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, there exists a permutation π of the estimated latent variables, denoted as $\hat{Z}_\pi$, such that each $\hat{Z}_{\pi(i)}$ is solely a function of a subset of the variables in $\{Z_i\} \cup \Psi_{Z_i}$.*



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.

(b) The corresponding Markov network $\mathcal{M}_Z$.

Figure 2: Illustrative example 2.

The proof is given in Appendix E. Roughly speaking, the proof leverages Theorems 1 and 2 to reason about the relationships among the true latent variables and the recovered ones, which imply that certain entries on the Jacobian matrix $\frac{\partial Z}{\partial \hat{Z}}$ must be zero. We show that these entries remain zero in the powers of $\frac{\partial Z}{\partial \hat{Z}}$, indicating that the same entries remain zero in $\left(\frac{\partial Z}{\partial \hat{Z}}\right)^{-1}$ (because the inverse $\left(\frac{\partial Z}{\partial \hat{Z}}\right)^{-1}$ can be written as a linear combination of the powers of $\frac{\partial Z}{\partial \hat{Z}}$) and thus $\frac{\partial \hat{Z}}{\partial Z}$, from which the identifiability result can be derived.

It is worth noting that in many cases, Theorem 3 already enables us to recover some of the latent variables up to a component-wise transformation.

Remark 1. *No matter how many neighbors each latent causal variable Z_i has, as long as each of its neighbors is not adjacent to at least one other neighbor in the Markov network $\mathcal{M}_Z$, then Z_i can be recovered up to a component-wise transformation.*

Even if the above case does not hold, Theorem 3 still shows how the estimated latent variables relate to the underlying causal variables in a specific, nontrivial way. Two examples are provided below.

Example 1. *Consider the Markov network $\mathcal{M}_Z$ corresponding to the DAG $\mathcal{G}_Z$ over Z in Figure 1. By Theorem 3 and suitable permutation of estimated latent variables $\hat{Z}$, we have: (a) $\hat{Z}_{\pi(1)}$ is solely a function of a subset of $\{Z_1, Z_2\}$, (b) $\hat{Z}_{\pi(2)}$ is solely a function of Z_2, (c) $\hat{Z}_{\pi(3)}$ is solely a function of a subset of $\{Z_2, Z_3, Z_4\}$, (d) $\hat{Z}_{\pi(4)}$ is solely a function of Z_4, and (e) $\hat{Z}_{\pi(5)}$ is solely a function of a subset of $\{Z_4, Z_5\}$. In this example, the latent causal variables Z_2 and Z_4 can be recovered up to component-wise transformation, while variables Z_1, Z_3, and Z_5 can be identified up to mixtures with certain neighbors in the Markov network.*

Example 2. *One may think that generally speaking, the more complex $\mathcal{G}_Z$, the more indeterminacies we have in the estimated latent variables (in the sense that each estimated latent variable receives contributions from more latent variables). In fact, this may not be the case. For instance, consider the underlying latent causal graph $\mathcal{G}_Z$ in Figure 2(a), which involves more variables and more edges and whose Markov network is shown in Figure 2(b). For every variable Z_i that is not the sink node, it has $\Psi_{Z_i} = \varnothing$ and thus can be recovered up to a component-wise transformation.*

Permutation of recovered latent variables. Theorems 2 and 3 involve certain permutation of the estimated latent variables $\hat{Z}$. Such an indeterminacy is common in the literature of causal discovery and representation learning tasks involving latent variables. In our case, since the function $v := g^{-1} \circ \hat{g}$ where $Z = v(\hat{Z})$ is invertible, there exists a permutation of the latent variables such that the corresponding Jacobian matrix J_v has nonzero diagonal entries (see Lemma 2 in Appendix A.1); such a permutation is what Theorems 2 and 3 refer to.

Connection with nonlinear ICA. It is worth noting that nonlinear ICA (with auxiliary variables) may be viewed as a special case of our result in this section. Specifically, if the true latent causal DAG $\mathcal{G}_Z$ is an empty graph, then the latent causal variables are independent, which reduce to the nonlinear ICA setting (Hyvärinen et al., 2019).

Furthermore, since traditional nonlinear ICA always has a valid solution (to produce nonlinear independent components) (Hyvärinen et al., 1999), one may wonder whether, in our setting, it is possible to always find nonlinear components as functions of X that are independent in each domain, as produced by recent methods for nonlinear ICA with auxiliary variables (Hyvärinen et al., 2019). As a corollary of Theorem 2, we show that the answer is no–there do not exist nonlinear components that are independent across domains if the true latent causal DAG $\mathcal{G}_Z$ is not an empty graph. The proof is provided in Appendix F.

Corollary 1 (Impossibility of finding independent components)**.** *Let the observations be sampled from the data generating process in Eq. (1). Suppose that Assumptions A1 and A2 from Theorem 1 hold, and that the true latent causal DAG $\mathcal{G}_Z$ is not an empty graph. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ with the components of $\hat{Z}$ being independent in each domain. Then, $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ cannot achieve Eq. (2).*

3.2. From Latent Markov Network to Latent Causal DAG

Now we have identified the Markov network up to an isomorphism, which characterizes conditional independence relations in the distribution. To build the connection between Markov network or conditional independence relations and causal structures, prior theory relies on the Markov and faithfulness assumptions. However, in real-world scenarios, the faithfulness assumption could be violated due to various reasons including path cancellations (Zhang & Spirtes, 2008; Uhler et al., 2013).

Since our goal is to generalize the identifiability theory as much as possible to fit practical applications, we introduce two relaxations of the faithfulness assumptions.

Assumption 1 (Single adjacency-faithfulness (SAF))**.** *Given a DAG $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ over the variable*

set Z, if two variables Z_i and Z_j are adjacent in $\mathcal{G}_Z$, then $Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,k\}}$.

Assumption 2 (Single unshielded-collider-faithfulness (SUCF) (Ng et al., 2021))**.** *Given a latent causal graph $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ over the variable set Z, let $Z_i \to Z_j \leftarrow Z_k$ be any unshielded collider in $\mathcal{G}_Z$, then $Z_i \not\perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{i,k\}}$.*

We propose SAF as a relaxtion of the Adjacency-faithfulness (Ramsey et al., 2012). The SUCF assumption is first introduced by Ng et al. (2021), which is strictly weaker than Orientation-faithfulness (Ramsey et al., 2012). Thus, both of them are strictly weaker than the faithfulness assumption, since the combination of Adjacency-faithfulness and Orientation-faithfulness is weaker than the faithfulness assumption (Zhang & Spirtes, 2008).

Interestingly, not only they are weaker variants of faithfulness, but we also prove that they are actually necessary and sufficient conditions, thus the weakest possible ones, to bridge conditional independence relations and causal structures. Specifically, we show that the recovered Markov network (e.g., in Theorem 2) is exactly the moralized graph of the true causal DAG if and only if the proposed variants of faithfulness hold. The proofs of Lemma 1 and Proposition 2 are shown in Appendix G.

Lemma 1. *Given a latent causal graph $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ with its Markov Network $\mathcal{M}_Z$, under Markov assumption, the undirected graph defined by $\mathcal{M}_Z$ is a subgraph of the moralized graph of the true causal DAG G.*

Proposition 2 (Moralized graph and Markov network)**.** *Given a causal DAG $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ with its Markov Network $\mathcal{M}_Z$, under Markov assumption, the undirected graph defined by $\mathcal{M}_Z$ is the moralized graph of the true causal DAG $\mathcal{G}_Z$ if and only if the SAF and SUCF assumptions are satisfied.*

It is worth noting that the connection between conditional independence relations and causal structures has been developed by (Loh & Bühlmann, 2014; Ng et al., 2021) in the linear case by leveraging the properties of the inverse covariance matrix; our results here focus on the nonparametric case and thus being able to serve the considered general settings for identifiability. Also note that the necessary and sufficient conditions may also be of independent interest for other causal discovery tasks exploring conditional independence relations in the nonparametric case.

Discussion on additional assumptions. We investigated how the sparsity constraint on the recovered graph over latent variables and sufficient change conditions on causal influences can be used to recover the latent variables and causal graph up to certain indeterminacies. Our framework is connected with previous ones in a spectrum of related studies (Varici et al., 2023; Ahuja et al., 2023; Buchholz et al., 2022; Squires et al., 2023; Brehmer et al., 2022; von Kügelgen et al., 2023; Brehmer et al., 2022; von Kügelgen et al., 2023; Zheng & Zhang, 2023; Zhang et al., 2023). For instance, the connection between conditional independence and cross-derivatives of the log density in both linear and nonlinear cases means our theorems directly apply to linear SEMs. Furthermore, our results do not require the mixing function to be sufficiently nonlinear, allowing them to encompass linear mixing processes as well.

At the same time, we may be able to leverage possible parametric constraints on the data generating process (or functions) or specific types of interventions. For instance, if we know that the changes happen to the linear causal mechanisms with Gaussian noises, this constraint can readily help reduce the search space and improve the identifiability. Moreover, since we only require the changing distribution, any type of interventions will be covered since any change to the conditional distribution is allowed. Given the additional information illustrated by experimental interventions (e.g., single-node interventions), alternative identifiability that might be particularly useful in certain tasks can be established. We hope this work can provide a helpful, bigger picture of causal representation learning in the general setting and further illustrates the necessity and connections of the different assumptions formulated in this line of works.

4. Change Encoding Network for Representation Learning

Thanks to the identifiability result, we now present two different practical implementations to recover the latent variables and their causal relations from observations from multiple domains. We build our method on the variational autoencoder (VAE) framework and can be easily extended to other models, such as normalizing flows.

We learn a deep latent generative model (decoder) $p(X|Z;\hat{\theta}^{(u)})$ and a variational approximation (encoder) $q(Z|X,u)$ of its true posterior $p(Z|X;\theta^{(u)})$ since the true posterior is usually intractable. To learn the model, we minimize the lower bound of the log-likelihood as

$$
\begin{aligned}
&\log p(X;\hat{\theta}^{(u)}) \\
&\quad = \log \int p(X|Z;\hat{\theta}^{(u)}) p(Z;\hat{\theta}^{(u)}) dZ \\
&\quad = \log \int \frac{q(Z|X,u)}{q(Z|X,u)} p(X|Z;\hat{\theta}^{(u)}) p(Z;\hat{\theta}^{(u)}) dZ \\
&\quad \geq -\mathrm{KL}(q(Z|X,u)||p(Z;\hat{\theta}^{(u)})) + \mathbb{E}_q[\log p(X|Z;\hat{\theta}^{(u)})] \\
&\quad = -\mathcal{L}_{\text{ELBO}}
\end{aligned}
$$

For the posterior $q(Z|X,u)$, we assume that it is a multivariate Gaussian or a Laplacian distribution, where the mean and variance are generated by the neural network encoder. As for $q(X|Z)$, we assume that it is a multivariate Gaussian

and the mean is the output of the decoder and the variance is a pre-defined value.

In practice, we can parameterize $p(X|Z;\hat{\theta}^{(u)})$ as the decoder which takes as input the latent representation Z and $q(Z|X,u)$ as an encoder which outputs the mean and scale of the posterior distribution. An essential difference between VAE (Kingma & Welling, 2013) and iVAE (Khemakhem et al., 2020a) is that our method allows the components of Z to be causally dependent and we are able to learn the components and causal relationships. And the key is the prior distribution $P(Z;\hat{\theta}^{(u)})$. Now we present two different implementations to capture the changes with a properly defined prior distribution.

4.1. Nonparametric Implementation of the Prior Distribution

To recover the relationships and latent variables Z, we build the normalizing flow to mimic the inverse of the latent SEM $Z_i = f_i(\text{PA}(Z_i), \epsilon_i)$ in Eq. (1). We first assume a causal ordering as $\hat{Z}_1, \ldots, \hat{Z}_n$. Then, for each component $\hat{Z}_i$, we consider the previous components $\{\hat{Z}_1, \ldots, \hat{Z}_{i-1}\}$ as potential parents of $\hat{Z}_i$ and we can select the true parents with the adjacency matrix $\hat{A}$, where $\hat{A}_{i,j}$ denotes that component $\hat{Z}_j$ contributes in the generation of $\hat{Z}_i$. If $\hat{A}_{i,j} = 0$, it means that $\hat{Z}_j$ will not contribute to the generation of $\hat{Z}_i$. Since $\theta^{(u)}$ governs the changes across domains, we use the observed domain index u to discover the changes. Then, we use the selected parents $\{\hat{A}_{i,1}\hat{Z}_1, \ldots, \hat{A}_{i,i-1}\hat{Z}_{i-1}\}$ and the domain label u to generate parameters of normalizing flow and apply the flow transformation on $\hat{Z}_i$ to turn it into $\hat{\epsilon}_i$. Specifically, we have

$$\hat{\epsilon}_i, \log \det_i = \text{Flow}(\hat{Z}_i; \text{NN}(\{\hat{A}_{i,j}\hat{Z}_j\}_{j=1}^{i-1}, u)),$$

where $\log \det_i$ is the log determinant of the conditional flow transformation on $\hat{Z}_i$ and NN represents a neural network.

To compute the prior distribution, we make an assumption on the noise term ϵ that it follows an independent prior distribution $p(\epsilon)$, such as a standard isotropic Gaussian or a Laplacian. Then according to the change-of-variable formula, the prior distribution of the dependent latents can be written as

$$\log p(\hat{Z};\theta^{(u)}) = \sum_{i=1}^{n} (\log p(\hat{\epsilon}_i) + \log \det_i).$$

Intuitively, to minimize the KL divergence loss between $p(Z;\hat{\theta}^{(u)})$ and $q(Z|X,u)$, the network has to learn the correct structure and the underlying latent variables; otherwise, it can be difficult to transform the dependent latent variables $\hat{Z}$ to a factorized prior distribution, e.g., $\mathcal{N}(0, \mathrm{I})$.

4.2. Parametric Implementation of the Prior Distribution

We can make parametric assumption on the latent causal process and facilitate the learning of true causal structure and components. Here, we consider the linear SEM and more complex SEMs can be generalized. Specifically, we assume that the true generation process of the latent Z is linear and only consists of scaling and shifting mechanisms:

$$Z = A(C^{(u)}Z) + S^{(u)}\epsilon + B^{(u)},$$

where $A \in [0,1]^{n\times n}$ is a causal adjacency matrix which can be permuted to be strictly lower-triangular, $C^{(u)} \in \mathbb{R}^{n\times n}$ and $S^{(u)} \in \mathbb{R}^{n\times 1}$ are underlying domain-specific scaling matrix and vector for domain u, respectively, $B^{(u)} \in \mathbb{R}^{n\times 1}$ is the underlying domain-specific shift vector, and ϵ is the independent noise.

To estimate the latent variables Z, the causal structure A, and capture the changes across domains, we introduce the learnable scaling $\hat{C} \in \mathbb{R}^{n\times n}, \hat{S} \in \mathbb{R}^{n\times 1}$and bias parameters $\hat{B} \in \mathbb{R}^{n\times 1}$ and pre-define a causal ordering as $\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n$. Then we have the matrix form as

$$\hat{\epsilon} = (\hat{Z} - \hat{B}^{(u)} - \hat{A}\hat{C}^{(u)}\hat{Z})/\hat{S}^{(u)}.$$

Note that the determinant of the strictly lower triangular matrix $\hat{C}$ is 0. Given a prior distribution of the noise term $p(\hat{\epsilon})$, and according to the change-of-variable rule, we then have the prior distribution for $\hat{Z}$ in parametric case as

$$\log p(\hat{Z};\hat{\theta}^{(u)}) = \sum_{i=1}^{n} (\log p(\hat{\epsilon}_i) - \log |\hat{S}_i^{(u)}|).$$

4.3. Full Objective

After we have properly defined the needed distributions $p(X|Z;\hat{\theta}^{(u)}), q(Z|X,u), p(Z;\hat{\theta}^{(u)})$, we can train our model to minimize the loss function $\mathcal{L}_{\text{ELBO}}$. However, without any further constraint, the powerful network may choose to use the fully connected causal graph during training. In other words, all lower-triangular elements of the estimated graph $\hat{A}$ is non-zero, which implies that each component $\hat{Z}_i$ is caused by all previous $i-1$ components. To exclude such unwanted solutions and encourage the model to learn the true causal structure among components of Z, we apply the ℓ_1 regularization on $\hat{A}$, i.e.,

$$\mathcal{L}_{\text{sparsity}} = \|\hat{A}\|_1.$$

It is worth noting that the sparsity regularization term above is an approximation of the sparsity constraint on the edges of the estimated Markov network specified in Theorems 2 and 3, since it is not straightforward to impose the latter constraint in a differentiable end-to-end training process.

A more sophisticated alternative is to impose sparsity constraint on $(I-\hat{A})^T\Omega^{-1}(I-\hat{A})$ where Ω is a randomly sampled positive diagonal matrix. Note that this corresponds to the formula of precision matrix whose nonzero entries represent the moral graph under certain conditions (Loh & Bühlmann, 2014) and we leave it for future investigation.

Finally, the full training objective is

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{sparsity}}.$$

After the model converges, the output of the encoder $\hat{Z}$ is our recovered latents from the observations in multiple domains and the revealed causal structure is in $\hat{A}$ which encapsulates the causal relationships across the components.

4.4. Simulations

To verify our theory and the proposed implementations, we run experiments on the simulated data because the ground truth causal adjacency matrix and the latent variables across domains are available for simulated data. Consequently, we consider following common causal structures (i) Y-structure with 4 variables, $Z_1 \to Z_3 \leftarrow Z_2, Z_3 \to Z_4$ and (ii) chain structure $Z_1 \to Z_2 \to Z_3 \to Z_4$. The noises are modulated with scaling random sampled from $\text{Unif}[0.5, 2]$ and shifts are sampled from $\text{Unif}[-2, 2]$. The scaling on the Z are also randomly sampled from $\text{Unif}[0.5, 2]$. In other words, the changes are modular. After generating Z, we feed the latent variables into multilayer perceptron (MLP) with orthogonal weights and LeakyReLU activations for invertibility. Specifically, we sample orthogonal matrix as the weights of the MLP layers. Since orthogonal matrix and LeakyReLU are invertible, the MLP function is also invertible.

We present the results in Figures 3 and 4. Each sub-figure consist of 4×4 panels and penal on i-th row and j-th column denotes the relationship between the estimated component $\hat{Z}_i$ with the true latent Z_j. We can see that under most cases, our model learns a strong one-to-one correspondence from the estimated components and the true components. For instance, the first column in Figure 3 show that $\hat{Z}_1$ is strongly correlated with the true components Z_1 while it is nearly independent from the true Z_2.

From the estimated $\hat{A}$, we find that our method is able to recover the true causal structure. For instance, on the Y-structure with $Z_1 \to Z_3 \leftarrow Z_2$ and $Z_3 \to Z_4$, our estimated model only keep the components $\hat{A}_{1,3}, \hat{A}_{2,3}, \hat{A}_{3,4}$ nonzero with the proposed sparsity regularization. The estimated causal graph is consistent with the true Y-structure causal graph. We can also see that the latent causal structure is also recovered from Figures 4 and 3. We observe that the learned $\hat{Z}_1$ is strongly correlated with the true Z_2 and is independent from the true Z_1, but correlated with the $\hat{Z}_3$ and $\hat{Z}_4$. These results align well with the true causal graph since Z_2 is independent from Z_1 while is the cause of Z_3 and Z_4.

The experiments support our theoretical result that the components and structure are identifiable up to certain indeterminacies. As for the results in Figure 3, we observe that our non-parametric method is still able to recover the true latent variables with Laplace noise.

5. Related Work

Causal representation learning aims to unearth causal latent variables and their relations from observed data. Despite its significance, the identifiability of the hidden generating process is known to be impossible without additional constraints, especially with only observational data. In the linear, non-Gaussian case, Silva et al. (2006) recover the Markov equivalence class, provided that each observed variable has a unique latent causal parent; Xie et al. (2020); Cai et al. (2019) estimate the latent variables and their relations assuming at least twice measured variables as latent ones, which has been further extended to learn the latent hierarchical structure (Xie et al., 2022). Moreover, Adams et al. (2021) provide theoretical results on the graphical conditions for identification. In the linear, Gaussian case, Huang et al. (2022) leverage rank deficiency of the observed sub-covariance matrix to estimate the latent hierarchical structure, while Dong et al. (2023) further extend the rank constraint to accommodate flexibly related latent and observed variables. In the discrete case, Kivva et al. (2021) identify the latent causal graph up to Markov equivalence by assuming a mixture model where the observed children sets of any pair of latent variables are different.

Given the challenge of identifiability on purely observational data, a different line of research leverage experiments by assuming the accessibility of various types of interventional data. Based on the single-node perfect intervention, Squires et al. (2023) leverage single-node interventions for the identifiability of linear causal model and linear mixing function; (Varici et al., 2023) incorporate for nonlinear causal model and linear mixing function; (Varici et al., 2023; Buchholz et al., 2023; Jiang & Aragam, 2023) provide the identifiability of the nonparametric causal model and linear mixing function; (Ahuja et al., 2023) further generalize the result to nonparametric causal model and polynomial mixing functions with additional constraints on the latent support; and (Brehmer et al., 2022; von Kügelgen et al., 2023; Jiang & Aragam, 2023) explore the nonparametric settings for both the causal model and mixing function. In addition to the single-node perfect interventions, Brehmer et al. (2022) introduced counterfactual pre- and post-intervention views; von Kügelgen et al. (2023) assume two distinct, paired interventions per node for multivariate causal models; Zhang et al. (2023) explore soft interventions on polynomial mixing functions; and Jiang & Aragam (2023) places specific structural restrictions on the latent causal graph.

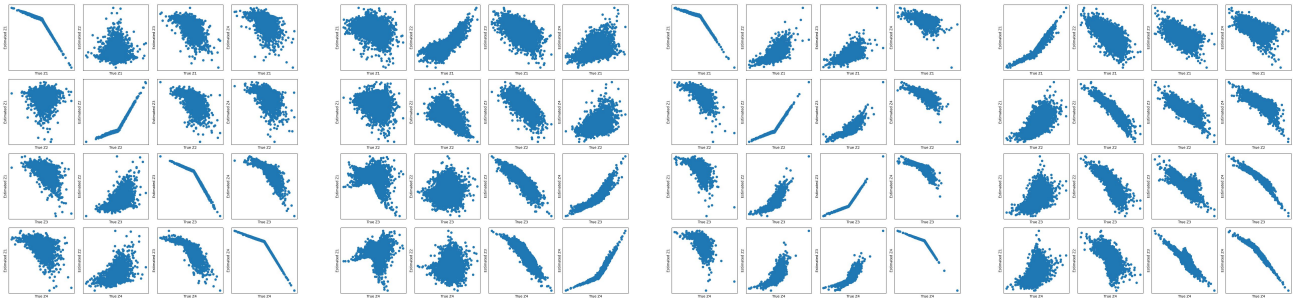


Figure 3: Recovered latent variables v.s. the true latent variables with Non-Parametric Approach. (a) Y-structure with Laplace noise. (b) Y-structure with Gaussian noise. (c) Chain structure with Laplace noise. (d) Chain structure with Gaussian noise. In each sub-figure, i-th row and j-th column depcits the relationship between the estimated $\hat{Z}_i$ and the true components Z_j.

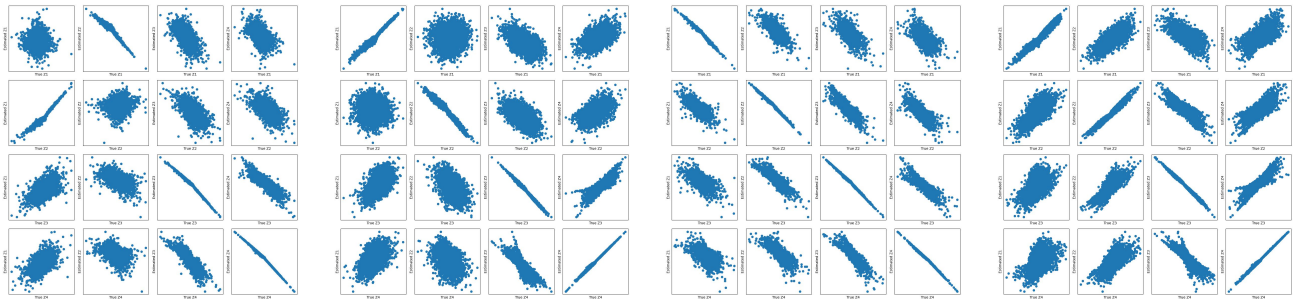


Figure 4: Recovered latent variables v.s. the true latent variables with Linear Parameterization Approach. The X-axis denotes the components of true latent variables Z and the Y-axis represent the components of estimated latent variables $\hat{Z}$. (a) Y-structure with Laplace noise. (b) Y-structure with Gaussian noise. (c) Chain structure with Laplace noise. (d) Chain structure with Gaussian noise.

Our study lies in the line of leveraging only observational data, and provides identifiability results in the general nonparametric settings on *both* the latent causal model and mixing function. Unlike prior works with observational data, we do not have any parametric assumptions or graphical restrictions; Compared to those relying on interventional data, our results naturally benefit from the heterogeneity of observational data (e.g., multi-domain data, nonstationary time series) and avoid additional experiments for interventions.

6. Conclusion and Discussions

We establish a set of new identifiability results to reveal latent causal variables and latent structures in the general nonparametric settings. Specifically, with sparsity regularization during estimation and sufficient changes in the causal influences, we demonstrate that the revealed latent variables and structures are related to the underlying causal model in a specific, nontrivial way. In contrast to recent works on the recovery of latent causal variables and structures, our results rely on purely observational data without graphical or parametric constraints. Our results offer insight into unveiling the latent causal process in one of the most universal settings. Experiments in various settings have been conducted to validate the theory. As future work, we will explore the scenario where only a subset of the causal relations change, which could be a challenge as well as a chance, and show up to what extent the underlying causal variables can be recovered with potentially weaker assumptions.

Acknowledgements

The authors would like to thank the anonymous reviewers for helpful comments and suggestions. The authors would also like to acknowledge the support from NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Florin Court Capital.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

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

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

Buchholz, S., Besserve, M., and Schölkopf, B. Function classes for identifiable nonlinear independent component analysis. *arXiv preprint arXiv:2208.06406*, 2022.

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

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

Comon, P. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

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

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

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

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

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

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*, 29: 3765–3773, 2016.

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

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

Hyvärinen, A., Cristescu, R., and Oja, E. A fast algorithm for estimating overcomplete ICA bases for image windows. In *Proc. Int. Joint Conf. on Neural Networks*, pp. 894–899, Washington, D.C., 1999.

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

Hyvärinen, A., Khemakhem, I., and Morioka, H. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2023.100844. URL https://www.sciencedirect.com/science/article/pii/S2666389923002234.

Jiang, Y. and Aragam, B. Learning nonparametric latent causal graphs with unknown interventions. *arXiv preprint arXiv:2306.02899*, 2023.

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

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, pp. 11455–11472. PMLR, 2022.

Lachapelle, S., López, P. R., Sharma, Y., Everett, K., Priol, R. L., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. *Conference on Causal Learning and Reasoning*, 2022.

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

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Ng, I., Zheng, Y., Zhang, J., and Zhang, K. Reliable causal discovery with improved exact search and weaker assumptions. In *Advances in Neural Information Processing Systems*, 2021.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Ramsey, J., Zhang, J., and Spirtes, P. L. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

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

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

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

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

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

Strang, G. *Linear Algebra and Its Applications*. Thomson, Brooks/Cole, Belmont, CA, 4th edition, 2006.

Strang, G. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 5th edition, 2016.

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

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pp. 436–463, 2013.

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

von Kügelgen, J., Besserve, M., Liang, W., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. *arXiv preprint arXiv:2306.00542*, 2023.

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

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

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

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

Zhang, J. and Spirtes, P. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

Zhang, J., Greenewald, K., Squires, C., Srivastava, A., Shanmugam, K., and Uhler, C. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2023.

Zheng, Y. and Zhang, K. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

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

Zheng, Y., Ng, I., Fan, Y., and Zhang, K. Generalized precision matrix for scalable estimation of nonparametric markov networks. *arXiv preprint arXiv:2305.11379*, 2023.

Supplementary Material

A. Proofs of Useful Lemmas

A.1. Proof of Lemma 2

The following lemma is a rather standard result in linear algebra (Strang, 2006; 2016), which has also been used in existing works in causal representation learning, such as Lachapelle et al. (2022). We provide the proof here for completeness.

Lemma 2. *For any invertible matrix A, there exists a permutation of its columns such that the diagonal entries of the resulting matrix are nonzero.*

Proof. Suppose by contradiction that there exists at least a zero diagonal entry for every column permutation. By Leibniz formula, we have

$$\det(A) = \sum_{\sigma \in \mathcal{S}_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)} \right),$$

where $\mathcal{S}_n$ denotes the set of n-permutations. Since there exists a zero diagonal entry for every permutation, we have

$$\prod_{i=1}^{n} A_{i,\sigma(i)} = 0, \quad \forall \sigma \in \mathcal{S}_n,$$

which implies $\det(A) = 0$ and that matrix A is not invertible. This is a contradiciton with the assumption that A is invertible. □

A.2. Proof of Lemma 3

Lemma 3. *Suppose matrix $A \in \mathbb{R}^{n \times n}$ contains a zero submatrix of order $(i+1) \times (n-i)$ for some $i \in [n-1]$. Then, matrix A is not invertible.*

Proof. By the given condition, there exist $n-i$ columns in matrix A of which $i+1$ rows are zero, i.e., at most $n-i-1$ rows are not zero. This implies that the $n-i$ column vectors span a subspace of dimension less than $n-i$, which thus are linearly dependent. Therefore, matrix A cannot be invertible. □

B. Proof of Proposition 1

Proposition 1. *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose the following assumptions hold:*

- *A1 (Smooth and positive density): The probability density function of latent causal variables, i.e., p_Z, is twice continuously differentiable and positive in $\mathbb{R}^n$.*
- *A2 (Sufficient changes): For each value of Z, there exist $2n+|\mathcal{M}_Z|+1$ values of θ, i.e., $\theta^{(u)}$ with $u = 0, \ldots, 2n+|\mathcal{M}_Z|$, such that the vectors $w(Z,u) - w(z,0)$ with $u = 1, \ldots, 2n+|\mathcal{M}_Z|$ are linearly independent, where vector $w(Z,u)$ is defined as follows:*

$$w(Z,u) = \left(\frac{\partial \log p(Z;\theta^{(u)})}{\partial Z_i} \right)_{i \in [n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i^2} \right)_{i \in [n]} \oplus \left(\frac{\partial^2 \log p(Z;\theta^{(u)})}{\partial Z_i \partial Z_j} \right)_{\{Z_i, Z_j\} \in \mathcal{E}(\mathcal{M}_Z),\, i<j}.$$

Suppose that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2). Then, for every pair of estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are **not adjacent in the Markov network** *$\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$, we have the following statements:*

(a) For each true latent causal variable Z_i, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k} \frac{\partial Z_i}{\partial \hat{Z}_l} = 0.$$

(b) For each pair of true latent causal variables Z_i and Z_j that are adjacent in the Markov network $\mathcal{M}_Z$, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_j}{\partial \hat{Z}_l} = 0.$$

Proof. For a matrix A, denote by $\operatorname{vol} A$ its volume, which is the product of its singular values. Note that $\operatorname{vol} A = \sqrt{\det AA^T}$ when A is of full row rank. In the change-of-variable formula, when the Jacobian is a rectangular matrix, the absolute determinant of the Jacobian can be replaced with the matrix volume (Ben-Israel, 1999; Gemici et al., 2016).

Since $X = g(Z)$ and $\hat{X} = \hat{g}(\hat{Z})$, by Eq. (2) and the change-of-variable formula, we have

$$p_{\hat{X}} = p_X \implies p_{\hat{g}(\hat{Z})} = p_{g(Z)} \implies p_{g^{-1}\circ\hat{g}(\hat{Z})} \operatorname{vol} J_{g^{-1}} = p_Z \operatorname{vol} J_{g^{-1}} \implies p_{v(\hat{Z})} = p_Z,$$

where $J_{g^{-1}}$ is the Jacobian matrix of g^{-1} and $v := g^{-1} \circ \hat{g}$ is a composition of diffeomorphisms (and hence also a diffeomorphism). Let J_v be the Jacobian matrix of v. The change-of-variable formula implies

$$\begin{aligned} p(\hat{Z};\hat{\theta})|\det J_{v^{-1}}| &= p(Z;\theta) \\ \log p(\hat{Z};\hat{\theta}) &= \log p(Z;\theta) + \log|\det J_v|. \end{aligned} \tag{6}$$

Suppose $\hat{Z}_k$ and $\hat{Z}_l$ are conditionally independent given $\hat{Z}_{[n]\setminus\{k,l\}}$ i.e., they are not adjacent in the Markov network over $\hat{Z}$. For each $\hat{\theta}$, by Lin (1997), we have

$$\frac{\partial^2 \log p(\hat{Z};\hat{\theta})}{\partial \hat{Z}_k \partial \hat{Z}_l} = 0. \tag{7}$$

To see what it implies, we find the first-order derivative of Eq. (6):

$$\frac{\partial \log p(\hat{Z};\hat{\theta})}{\partial \hat{Z}_k} = \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial Z_i}{\partial \hat{Z}_k} + \frac{\partial \log|\det J_v|}{\partial \hat{Z}_k}.$$

Let

$$\eta(\theta) := \log p(Z;\theta), \quad \eta_i'(\theta) := \frac{\partial \log p(Z;\theta)}{\partial Z_i}, \quad \eta_{ij}''(\theta) := \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}, \quad h_{i,l}' := \frac{\partial Z_i}{\partial \hat{Z}_l}, \quad \text{and} \quad h_{i,kl}'' := \frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l}.$$

We then derive the second-order derivative w.r.t. $\hat{Z}_k$ and $\hat{Z}_l$ and apply Eq. (7):

$$\begin{aligned} 0 &= \sum_{j=1}^n \sum_{i=1}^n \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}\frac{\partial Z_j}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} + \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l} + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l}. \\ &= \sum_{i=1}^n \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i^2}\frac{\partial Z_i}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} + \sum_{j=1}^n \sum_{i:\{Z_j,Z_i\}\in\mathcal{E}(\mathcal{M}_Z)} \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}\frac{\partial Z_j}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} \\ &\quad + \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l} + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l} \end{aligned} \tag{8}$$

$$= \sum_{i=1}^n \eta_{ii}''(\theta) h_{i,l}' h_{i,k}' + \sum_{j=1}^n \sum_{i:\{Z_j,Z_i\}\in\mathcal{E}(\mathcal{M}_Z)} \eta_{ij}''(\theta) h_{j,l}' h_{i,k}' + \sum_{i=1}^n \eta_i'(\theta) h_{i,kl}'' + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l}. \tag{9}$$

Recall that $\mathcal{E}(\mathcal{M}_Z)$ denotes the set of edges in the Markov network over Z. In the equation above, we made use of the fact that if Z_i and Z_j are not adjacent in the Markov network, then $\frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j} = 0$ by Lin (1997).

By Assumption A2, consider the $2n + |\mathcal{M}_Z| + 1$ values of θ, i.e., $\theta^{(u)}$ with $u = 0, \ldots, 2n + |\mathcal{M}_Z|$, such that Eq. (9) hold. Then, we have $2n + |\mathcal{M}_Z| + 1$ such equations. Subtracting each equation corresponding to $\theta^{(u)}, u = 1, \ldots, 2n + |\mathcal{M}_Z|$ with the equation corresponding to $\theta^{(0)}$ results in $2n + |\mathcal{M}_Z|$ equations:

$$0 = \sum_{i=1}^n (\eta_{ii}''(\theta^{(u)}) - \eta_{ii}''(\theta^{(0)})) h_{i,l}' h_{i,k}' + \sum_{j=1}^n \sum_{i:\{Z_j,Z_i\}\in\mathcal{E}(\mathcal{M}_Z)} (\eta_{ij}''(\theta^{(u)}) - \eta_{ij}''(\theta^{(0)})) h_{j,l}' h_{i,k}' + \sum_{i=1}^n (\eta_i'(\theta^{(u)}) - \eta_i'(\theta^{(0)})) h_{i,kl}'', \tag{10}$$

where $u = 1, \ldots, 2n + |\mathcal{M}_Z|$. Since p_Z is twice continuously differentiable, we have

$$\eta''_{ij}(\theta^{(u)}) - \eta''_{ij}(\theta^{(0)}) = \eta''_{ji}(\theta^{(u)}) - \eta''_{ji}(\theta^{(0)}),$$

and therefore Eq. (10) can be written as

$$0 = \sum_{i=1}^{n} (\eta''_{ii}(\theta^{(u)}) - \eta''_{ii}(\theta^{(0)}))h'_{i,l}h'_{i,k} + \sum_{\substack{i,j:\\ i<j,\\ \{Z_i,Z_j\}\in\mathcal{E}(\mathcal{M}_Z)}} (\eta''_{ij}(\theta^{(u)}) - \eta''_{ij}(\theta^{(0)}))(h'_{j,l}h'_{i,k} + h'_{i,l}h'_{j,k}) + \sum_{i=1}^{n} (\eta'_i(\theta^{(u)}) - \eta'_i(\theta^{(0)}))h''_{i,kl}.$$

Consider the vectors formed by collecting the corresponding coefficients in the equation above where $u = 1, \ldots, 2n+|\mathcal{M}_Z|$. By Assumption A2, these $2n + |\mathcal{M}_Z|$ vectors are linearly independent. Thus, for any i and j such that $\{Z_i, Z_j\} \in \mathcal{E}(\mathcal{M}_Z)$, we have the following equations:

$$h'_{i,k}h'_{i,l} = 0, \tag{11}$$
$$h'_{i,k}h'_{j,l} + h'_{j,k}h'_{i,l} = 0, \tag{12}$$
$$h''_{i,kl} = 0.$$

It remains to show $h'_{i,k}h'_{j,l} = 0$. Suppose by contradiction that

$$h'_{i,k}h'_{j,l} \neq 0, \tag{13}$$

which implies $h'_{i,k} \neq 0$. By Eq. (11), we have $h'_{i,l} = 0$, which, by plugging into Eq. (12), indicates $h'_{i,k}h'_{j,l} = 0$. This is a contradiction with Eq. (13). Thus, we must have $h'_{i,k}h'_{j,l} = 0$. □

C. Proof of Theorem 2

Theorem 2 (Identifiability of latent Markov network)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, the recovered latent Markov network $\mathcal{M}_{\hat{Z}}$ is isomorphic to the true latent Markov network $\mathcal{M}_Z$.*

Proof. Let $v := g^{-1} \circ \hat{g}$, i.e., $Z = v(\hat{Z})$. Note that v is a composition of diffeomorphisms, and hence also a diffeomorphism. Consider a specific value of $\hat{Z}$, say $\hat{z}$. Since v is diffeomorphism, by Lemma 2, there exists a permutation π such that the diagonal entries of the corresponding Jacobian matrix (whose columns are permuted according to π) evaluated at $\hat{Z} = \hat{z}$ are nonzero, i.e.,

$$\left.\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}}\right|_{\hat{Z}=\hat{z}} \neq 0, \quad i = 1, \ldots, n. \tag{14}$$

Suppose that Z_i and Z_j are adjacent in the Markov network $\mathcal{M}_Z$ over Z, but $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$ are not adjacent in the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. By Proposition 1, we have

$$\left.\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}}\right|_{\hat{Z}=\hat{z}} \left.\frac{\partial Z_j}{\partial \hat{Z}_{\pi(j)}}\right|_{\hat{Z}=\hat{z}} = 0,$$

which is clearly a contradiction with Eq. (14).

Thus, we have shown by contradiction the following lemma.

Lemma 4. *If Z_i and Z_j are adjacent in the Markov network $\mathcal{M}_Z$ over Z, then $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$ are adjacent in the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$.*

The lemma above indicates

$$|\mathcal{M}_{\hat{Z}}| \geq |\mathcal{M}_Z|. \tag{15}$$

Also, note that the true model (g, f, p_Z, Θ) is one of the solutions that achieves Eq. (2). Since the recovered latent Markov network $\mathcal{M}_{\hat{Z}}$ has the minimal number of edges among the solutions that achieve Eq. (2), we have $|\mathcal{M}_{\hat{Z}}| \leq |\mathcal{M}_Z|$, which, with Eq. (15), implies $|\mathcal{M}_{\hat{Z}}| = |\mathcal{M}_Z|$.

By Lemma 4 and $|\mathcal{M}_{\hat{Z}}| = |\mathcal{M}_Z|$, we conclude that Z_i and Z_j are adjacent in $\mathcal{M}_Z$ if and only if $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$ are adjacent in $\mathcal{M}_{\hat{Z}}$. That is, $\mathcal{M}_Z$ and $\mathcal{M}_{\hat{Z}}$ are isomorphic. □

D. Proof of Theorem 1

Theorem 1 (Relations among true and recovered latent causal variables)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, for every pair of estimated latent variables $\hat{Z}_k$ and $\hat{Z}_l$ that are* **not adjacent in the Markov network** *$\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$, we have the following statements:*

(a) Each true latent causal variable Z_i is a function of at most one of $\hat{Z}_k$ and $\hat{Z}_l$.

(b) For each pair of true latent causal variables Z_i and Z_j that are adjacent in the Markov network $\mathcal{M}_Z$ over Z, at most one of them is a function of $\hat{Z}_k$ or $\hat{Z}_l$.

Proof. We first prove Statement (a). By Proposition 1, for every value of Z, we have

$$\frac{\partial Z_i}{\partial \hat{Z}_k}\frac{\partial Z_i}{\partial \hat{Z}_l} = 0.$$

Therefore, it suffices to prove that if $\frac{\partial Z_i}{\partial \hat{Z}_k} \neq 0$ for some value of $\hat{Z}$, then $\frac{\partial Z_i}{\partial \hat{Z}_l} = 0$ for all values of $\hat{Z}$. That is, these nonzero entries cannot switch positions.

By Theorem 2, there exists a permutation π of the estimated variables, denoted as $\hat{Z}_\pi$, such that the Markov network $\mathcal{M}_{\hat{Z}_\pi}$ is identical to $\mathcal{M}_Z$. Let $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(k)}$ be two estimated latent variables that are not adjacent in the Markov network $\mathcal{M}_{\hat{Z}_\pi}$. Now consider variable Z_i. Suppose by contradiction that the nonzero entries switch positions, i.e., there exist two values of $\hat{Z}$, say $\hat{z}^{(1)}$ and $\hat{z}^{(2)}$, such that

$$\left.\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}}\right|_{\hat{Z}=\hat{z}^{(1)}} \neq 0 \tag{16}$$

and

$$\left.\frac{\partial Z_i}{\partial \hat{Z}_{\pi(k)}}\right|_{\hat{Z}=\hat{z}^{(2)}} \neq 0, \tag{17}$$

Let N_{Z_i} be a set containing the indices of the neighbors of Z_i in $\mathcal{M}_Z$, and $N_{\hat{Z}_{\pi(i)}}$ be a set containing the indices of the neighbors of $Z_{\pi(i)}$ in $\mathcal{M}_{\hat{Z}_\pi}$. Similarly, let S_{Z_i} be a set containing the indices of the variables that are not adjacent to Z_i in $\mathcal{M}_Z$, and $S_{\hat{Z}_{\pi(i)}}$ be a set containing the indices of the variables that are not adjacent to of $Z_{\pi(i)}$ in $\mathcal{M}_{\hat{Z}_\pi}$. By definition, we have

$$N_{Z_i} \cup S_{Z_i} \cup \{i\} = [n], \tag{18}$$

which are pairwise disjoint.

Since $\mathcal{M}_{\hat{Z}_\pi}$ and $\mathcal{M}_Z$ are identical, we have $N_{Z_i} = N_{\hat{Z}_{\pi(i)}}$ and $S_{Z_i} = S_{\hat{Z}_{\pi(i)}}$. Now define the following function

$$\phi(\hat{Z}) = \sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(i)}}\right)^2 - \sum_{l \in S_{Z_i}} \sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\right)^2.$$

Plugging in $\hat{Z} = \hat{z}^{(1)}$, for $l \in S_{Z_i} = S_{\hat{Z}_{\pi(i)}}$ and $j \in N_{Z_i} \cup \{i\}$, Proposition 1 implies

$$\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(1)}} \frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\bigg|_{\hat{Z}=\hat{z}^{(1)}} = 0,$$

which, with Eq. (16), indicates

$$\frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\bigg|_{\hat{Z}=\hat{z}^{(1)}} = 0.$$

Substituting the above equation and Eq. (16) into function ϕ, we have

$$\phi(\hat{Z})|_{\hat{Z}=\hat{z}^{(1)}} = \sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(1)}} \right)^2 \geq \left(\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(1)}} \right)^2 > 0. \tag{19}$$

Now plug in $\hat{Z} = \hat{z}^{(2)}$. For $j \in N_{Z_i} \cup \{i\}$, Proposition 1 implies

$$\frac{\partial Z_j}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(2)}} \frac{\partial Z_i}{\partial \hat{Z}_{\pi(k)}}\bigg|_{\hat{Z}=\hat{z}^{(2)}} = 0,$$

which, with Eq. (17), indicates

$$\frac{\partial Z_j}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(2)}} = 0.$$

Substituting the above equation and Eq. (17) into function ϕ, we have

$$\phi(\hat{Z})|_{\hat{Z}=\hat{z}^{(2)}} = -\sum_{l \in S_{Z_i}} \sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\bigg|_{\hat{Z}=\hat{z}^{(2)}} \right)^2 \leq -\left(\frac{\partial Z_i}{\partial \hat{Z}_{\pi(k)}}\bigg|_{\hat{Z}=\hat{z}^{(2)}} \right)^2 < 0. \tag{20}$$

Since function ϕ is continuous (because all the partial derivatives involved are continuous) and its domain is a connected set, by applying Intermediate Value Theorem with Eqs. (19) and (20), there exists a value of $\hat{Z}$ in the domain, say $\hat{z}^{(3)}$, such that

$$\phi(\hat{Z})|_{\hat{Z}=\hat{z}^{(3)}} = 0,$$

which, by plugging the definition of function ϕ, implies

$$\sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(i)}}\bigg|_{\hat{Z}=\hat{z}^{(3)}} \right)^2 = \sum_{l \in S_{Z_i}} \sum_{j \in N_{Z_i} \cup \{i\}} \left(\frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\bigg|_{\hat{Z}=\hat{z}^{(3)}} \right)^2.$$

Note that if any of the terms in the summation on the left hand side (LHS) is nonzero, then, by Proposition 1, all terms in the summation on the right hand side (RHS) must be zero; in this case, LHS is nonzero but RHS equals zero, which is a contradiction. Similarly, if any of the terms in the summation on the RHS is nonzero, then, by Proposition 1, all terms in the summation on the LHS must be zero; in this case, RHS is nonzero but LHS equals zero, which is a contradiction. This implies that all terms in the summation on both LHS and RHS must be zero, i.e.,

$$\frac{\partial Z_j}{\partial \hat{Z}_{\pi(l)}}\bigg|_{\hat{Z}=\hat{z}^{(3)}} = 0 \quad \text{for} \quad j \in N_{Z_i} \cup \{i\}, l \in S_{Z_i} \cup \{i\}.$$

Since $|S_{Z_i} \cup \{i\}| = n - |N_{Z_i}|$ by Eq. (18), Lemma 3 indicates that the matrix $\frac{\partial Z}{\partial \hat{Z}_\pi}\big|_{\hat{Z}=\hat{z}^{(3)}}$ is not invertible. Thus, the (Jacobian) matrix $\frac{\partial Z}{\partial \hat{Z}}\big|_{\hat{Z}=\hat{z}^{(3)}}$ is also not invertible, which is a contradiction because the mapping from $\hat{Z}$ to Z is a diffeomorphism (specifically a a composition of diffeomorphisms).

Therefore, we have just proved Statement (a) by contradiction. Similar reasoning can be straightforwardly applied to prove Statement (b) and is omitted here. □

E. Proof of Theorem 3

We first state the following lemma that is used to prove Theorem 3. The proof is a straightforward consequence of Cayley–Hamilton theorem and is omitted here.

Lemma 5. *Let A be an $n \times n$ invertible matrix. Then, it can be expressed as a linear combination of the powers of A, i.e.,*

$$A^{-1} = \sum_{k=0}^{n-1} c_k A^k$$

for some appropriate choice of coefficients $c_0, c_1, \ldots, c_{n-1}$.

Now consider the Markov network $\mathcal{M}_Z$ over variables Z. With a slight abuse of notation, let N_{Z_i} be the set of neighbors of Z_i in $\mathcal{M}_Z$. The following result relates a matrix to its inverse, given that the matrix satisfies certain property defined by $\mathcal{M}_Z$.

Proposition 3. *Consider Markov network $\mathcal{M}_Z$ over Z. Let N_{Z_i} be the set of neighbors of Z_i in $\mathcal{M}_Z$, and A be an $n \times n$ invertible matrix. For each $i \neq j$ where Z_j is not adjacent to some nodes in $\{Z_i\} \cup N_{Z_i}$, suppose $A_{ij} = 0$. Then, $A^{-1}_{ij} = 0$.*

Proof. By Lemma 5, A^{-1} can be expressed as linear combination of the powers of A. Therefore, it suffices to prove that each matrix power A^k satisfies the following property: $A^k_{ij} = 0$ for each $i \neq j$ where Z_j is not adjacent to some nodes in $\{Z_i\} \cup N_{Z_i}$. We proceed with mathematical induction on k. By definition, the property holds in the base case where $k = 1$.

Now suppose that the property holds for A^k. We prove by contradiction that the property holds for A^{k+1}. Suppose by contradiction that $A^{k+1}_{ij} \neq 0$ for some $i \neq j$ where Z_j is not adjacent to some nodes in $\{Z_i\} \cup N_{Z_i}$. This implies that one of the following cases holds:

- Case (a): Z_j is not adjacent to Z_i in $\mathcal{M}_Z$.
- Case (b): There exists $Z_l \in N_{Z_i} \setminus \{Z_j\}$ such that Z_j and Z_l are not adjacent in $\mathcal{M}_Z$.

Since $A^{k+1}_{ij} = \sum_{r=0}^{n} A^k_{ir} A_{rj}$, the assumption $A^{k+1}_{ij} \neq 0$ implies that there must exist m such that $A^k_{im} A_{mj} \neq 0$, i.e., $A^k_{im} \neq 0$ and $A_{mj} \neq 0$. Since both A^k and A satisfy the property, this indicates (i) Z_m is adjacent to Z_i and all nodes in $N_{Z_i} \setminus \{Z_m\}$, and (ii) Z_j is adjacent to Z_m and all nodes in $N_{Z_m} \setminus \{Z_j\}$. We consider the following cases:

- Case of $m = l$: By (ii), Z_j is adjacent to Z_l, which contradicts Case (b) above. Also, we know that Z_l is adjacent to Z_i by (i), which indicates that Z_i is adjacent to Z_j, contradicting Case (a) above.
- Case of $m \neq l$: By (i) and (ii), Z_m is adjacent to Z_i and Z_j is adjacent to Z_m, implying that Z_i and Z_j are adjacent, which is contradictory with Case (a) above. Furthermore, since Z_l is a neighbor of Z_i, we know that Z_m and Z_l are adjacent by (i). Also, by (ii), Z_j is adjacent to Z_l, which contradicts Case (b) above.

In each of the cases above, there is a contradiction. □

We are now ready to prove the following result.

Theorem 3 (Identifiability of latent causal variables)**.** *Let the observations be sampled from the data generating process in Eq. (1), and $\mathcal{M}_Z$ be the Markov network over Z. Let N_{Z_i} be the set of neighbors of variable Z_i in $\mathcal{M}_Z$. Suppose that Assumptions A1 and A2 from Theorem 1 hold. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ to achieve Eq. (2) with the minimal number of edges of the Markov network $\mathcal{M}_{\hat{Z}}$ over $\hat{Z}$. Then, there exists a permutation π of the estimated latent variables, denoted as $\hat{Z}_\pi$, such that each $\hat{Z}_{\pi(i)}$ is solely a function of a subset of the variables in $\{Z_i\} \cup \Psi_{Z_i}$.*

Proof. We first prove a simpler case: there exists a permutation π of the estimated latent variables, denoted as $\hat{Z}_\pi$, such that Z_i is solely a function of $\hat{Z}_{\pi(i)}$ and a subset of the variables in $\{\hat{Z}_{\pi(r)} \mid Z_r \in \Psi_{Z_i}\}$.

By Theorem 2 and its proof, there exists a permutation π of the estimated variables, denoted as $\hat{Z}_\pi$, such that the Markov network $\mathcal{M}_{\hat{Z}_\pi}$ over $\hat{Z}_\pi$ is identical to $\mathcal{M}_Z$, and that

$$\frac{\partial Z_i}{\partial \hat{Z}_{\pi(i)}} \neq 0, \quad i = 1, \ldots, n.$$

Clearly, each variable Z_i is a function of $\hat{Z}_{\pi(i)}$.

We first show that if Z_j is not adjacent to Z_i in $\mathcal{M}_Z$, then Z_i cannot be a function of $\hat{Z}_{\pi(j)}$. Since Z_i and Z_j are not adjacent in $\mathcal{M}_Z$, we know that $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$ are not adjacent in $\mathcal{M}_{\hat{Z}_\pi}$. By Theorem 1, Z_i is a function of at most one of $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$, which implies that Z_i cannot be a function of $\hat{Z}_{\pi(j)}$, because we have shown that Z_i is a function of $\hat{Z}_{\pi(i)}$.

To refine further, now suppose that Z_j is adjacent to Z_i, but not adjacent to some $Z_k \in N_{Z_i} \setminus \{Z_j\}$. Since $\mathcal{M}_Z$ and $\mathcal{M}_{\hat{Z}_\pi}$ are identical, $\hat{Z}_{\pi(j)}$ is also not adjacent to $\hat{Z}_{\pi(k)}$ in $\mathcal{M}_{\hat{Z}_\pi}$. Since Z_i and Z_k are adjacent in $\mathcal{M}_Z$, by Theorem 1, at most one of them is a function of $\hat{Z}_{\pi(j)}$ or $\hat{Z}_{\pi(k)}$. This implies that Z_i cannot be a function of $\hat{Z}_{\pi(j)}$, because we have shown that Z_k is a function of $\hat{Z}_{\pi(k)}$.

Therefore, we have just shown that Z_i is solely a function of $\hat{Z}_{\pi(i)}$ and a subset of the variables in $\{\hat{Z}_{\pi(r)} \mid Z_r \in \Psi_{Z_i}\}$. Now consider variable $Z_l \notin \{Z_i\} \cup \Psi_{Z_i}$. Since Z_i is not a function of $\hat{Z}_{\pi(l)}$, we have

$$\left(\frac{\partial Z}{\partial \hat{Z}_\pi}\right)_{il} = \frac{\partial Z_i}{\partial \hat{Z}_{\pi(l)}} = 0.$$

By applying Proposition 3 with matrix $\frac{\partial Z}{\partial \hat{Z}_\pi}$, we have

$$\left(\frac{\partial Z}{\partial \hat{Z}_\pi}\right)^{-1}_{il} = 0,$$

which, by Inverse Function Theorem, implies

$$\frac{\partial \hat{Z}_{\pi(i)}}{\partial Z_l} = \left(\frac{\partial \hat{Z}_\pi}{\partial Z}\right)_{il} = \left(\frac{\partial Z}{\partial \hat{Z}_\pi}\right)^{-1}_{il} = 0.$$

Since the above equation holds for all values of Z, we conclude that $\hat{Z}_{\pi(i)}$ cannot be a function of Z_l. □

F. Proof of Corollary 1

Corollary 1 (Impossibility of finding independent components)**.** *Let the observations be sampled from the data generating process in Eq. (1). Suppose that Assumptions A1 and A2 from Theorem 1 hold, and that the true latent causal DAG $\mathcal{G}_Z$ is not an empty graph. Suppose also that we learn $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ with the components of $\hat{Z}$ being independent in each domain. Then, $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ cannot achieve Eq. (2).*

Proof. Suppose by contradiction that $(\hat{g}, \hat{f}, p_{\hat{Z}}, \hat{\Theta})$ achieves Eq. (2). By assumption, the components of $\hat{Z}$ are independent in each domain, indicating that the Markov network $\mathcal{M}_{\hat{Z}}$ is an empty graph. By the same reasoning in the proof of Theorem 2 (specifically Lemma 4), there exists a permutation π such that: if Z_i and Z_j are adjacent in $\mathcal{M}_Z$, then $\hat{Z}_{\pi(i)}$ and $\hat{Z}_{\pi(j)}$ are adjacent in $\mathcal{M}_{\hat{Z}}$. Since $\mathcal{M}_{\hat{Z}}$ is an empty graph, this implies that $\mathcal{M}_Z$ is also an empty graph, which is contradictory with the assumption that $\mathcal{G}_Z$ is not an empty graph. □

G. Proof of Lemma 1 and Proposition 2

Lemma 1. *Given a latent causal graph $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ with its Markov Network $\mathcal{M}_Z$, under Markov assumption, the undirected graph defined by $\mathcal{M}_Z$ is a subgraph of the moralized graph of the true causal DAG G.*

Proof. Let Z_j and Z_k, $j \neq k$ be two variables that are not adjacent in the moralized graph of $\mathcal{G}_Z$. Then it suffices to show that $\{Z_j, Z_k\} \notin \mathcal{E}(\mathcal{M}_Z)$. Because they are not adjacent in the moralized graph of $\mathcal{G}_Z$, they must not be adjacent in $\mathcal{G}_Z$ and must not share a common child in $\mathcal{G}_Z$. Thus, Z_j and Z_k are d-separated conditioning on $Z_{[n]\setminus\{j,k\}}$, which implies the conditional independence $Z_j \perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{j,k\}}$ based on the Markov assumption on $(\mathcal{G}_Z, P_{Z;\theta})$. Then we have $\{Z_j, Z_k\} \notin \mathcal{E}(\mathcal{M}_Z)$. □

Proposition 2 (Moralized graph and Markov network)**.** *Given a causal DAG $\mathcal{G}_Z$ and distribution $P_{Z;\theta}$ with its Markov Network $\mathcal{M}_Z$, under Markov assumption, the undirected graph defined by $\mathcal{M}_Z$ is the moralized graph of the true causal DAG $\mathcal{G}_Z$ if and only if the SAF and SUCF assumptions are satisfied.*

Proof. We prove both directions as follows.

Sufficient condition. We prove it by contradiction. Suppose that the structure defined by $\mathcal{M}_Z$ is not equivalent to the moralized graph of $\mathcal{G}_Z$. Then, according to Lemma 1, there exists a pair of variables Z_j and Z_k, $j \neq k$ that are adjacent in the moralized graph but $\{Z_j, Z_k\} \notin \mathcal{E}(\mathcal{M}_Z)$. Thus, we have $Z_j \perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{j,k\}}$. Then we consider the following two cases:

- If variables Z_j and Z_k correspond to a pair of neighbors in $\mathcal{G}_Z$, then they are adjacent. Together with the conditional independence relation $Z_j \perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{j,k\}}$, this implies that the SAF assumption is violated.
- If variables Z_j and Z_k correspond to a pair of non-adjacent spouses in $\mathcal{G}_Z$. Then they have an unshielded collider, indicating that the SUCF assumption is violated.

Necessary condition. We prove it by contradiction. Suppose SUCF or SAF is violated, we have the following two cases:

- Suppose SUCF is violated, i.e., there exists an unshielded collider $Z_j \to Z_i \leftarrow Z_k$ in the DAG $\mathcal{G}_Z$ such that $Z_j \perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{j,k\}}$. This conditional independence relation indicates that $\{Z_j, Z_k\} \notin \mathcal{E}(\mathcal{M}_Z)$. Since Z_j and Z_k are spouses, there exists an edge between them in the moralized graph of $\mathcal{G}_Z$, but is not contained in the structure defined by $\mathcal{M}_Z$, showing that they are not the same.
- Suppose SAF is violated, i.e., there exists a pair of neighbors Z_j and Z_k in the DAG $\mathcal{G}_Z$ such that $Z_j \perp\!\!\!\perp Z_k \mid Z_{[n]\setminus\{j,k\}}$. This conditional independence relation indicates that $\{Z_j, Z_k\} \notin \mathcal{E}(\mathcal{M}_Z)$. Because Z_j and Z_k are adjacent in $\mathcal{G}_Z$, clearly they are also adjacent in the moralized graph of $\mathcal{G}_Z$. However, the edge between them is not contained in the structure defined by $\mathcal{M}_Z$, showing that they are not the same.

Thus, when SUCF or SAF is violated, the structure defined by $\mathcal{M}_Z$ is the moralized graph of the true DAG $\mathcal{G}_Z$. □